MST121 CB D

The Open University

A first level interdisciplinary course

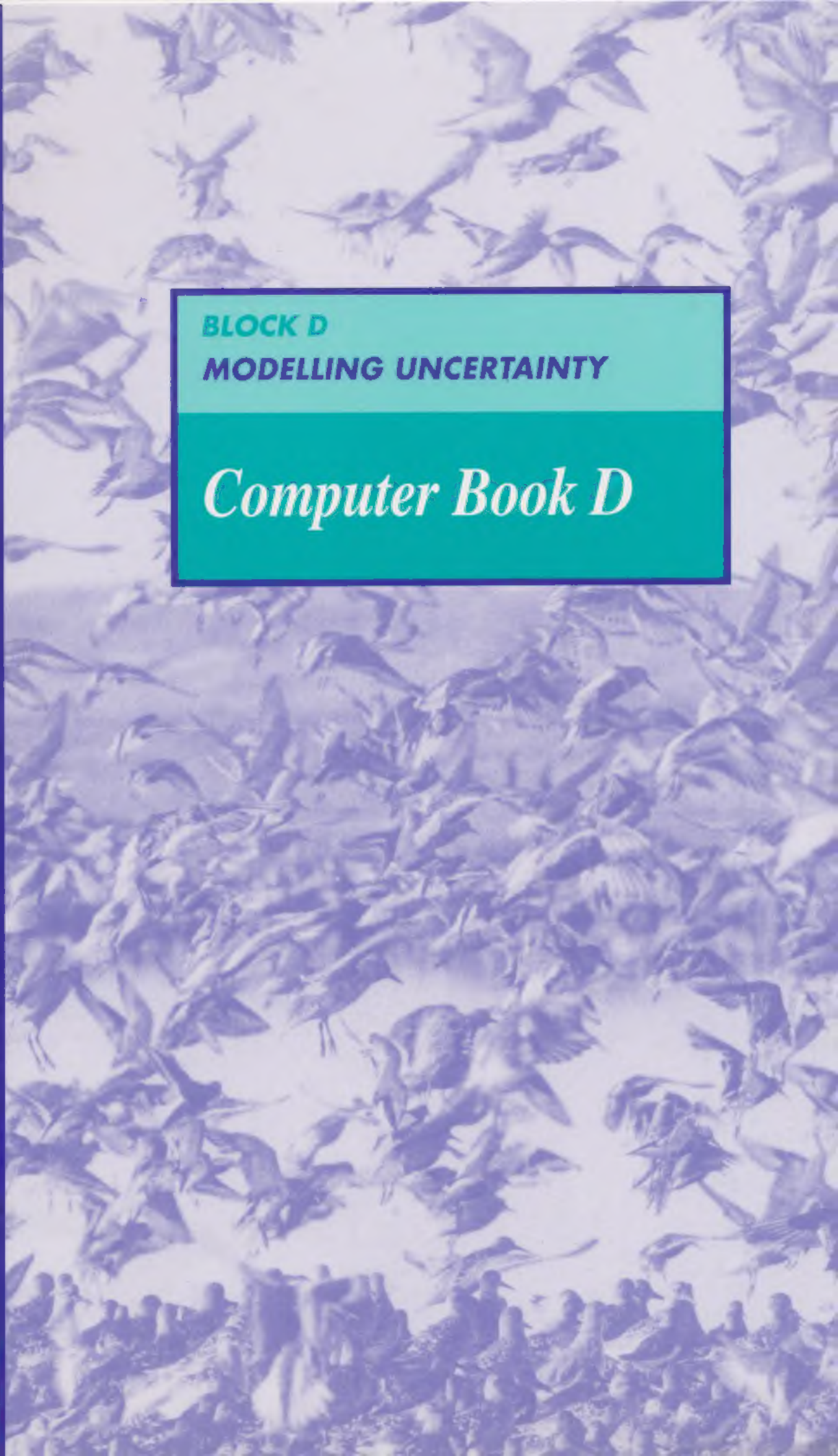
Using Mathematics

COMPUTER BOOK D

BLOCK D

MODELLING UNCERTAINTY

Computer Book D



MST121 CB D

The Open University

A first level interdisciplinary course

Using Mathematics

COMPUTER BOOK D

BLOCK D

MODELLING UNCERTAINTY

Computer Book D

Prepared by the course team

About this course

This computer book forms part of the course MST121 *Using Mathematics*. This course and the courses MU120 *Open Mathematics* and MS221 *Exploring Mathematics* provide a flexible means of entry to university-level mathematics. Further details may be obtained from the address below.

MST121 uses the software program Mathcad (MathSoft, Inc.) and other software to investigate mathematical and statistical concepts and as a tool in problem solving. This software is provided as part of the course.

This publication forms part of an Open University course. Details of this and other Open University courses can be obtained from the Student Registration and Enquiry Service, The Open University, PO Box 197, Milton Keynes, MK7 6BJ, United Kingdom: tel. +44 (0)870 333 4340, e-mail general-enquiries@open.ac.uk

Alternatively, you may visit the Open University website at http://www.open.ac.uk where you can learn more about the wide range of courses and packs offered at all levels by The Open University.

To purchase a selection of Open University course materials, visit the webshop at www.ouw.co.uk, or contact Open University Worldwide, Michael Young Building, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom, for a brochure: tel. +44 (0)1908 858785, fax +44 (0)1908 858787, e-mail ouwenq@open.ac.uk

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 1997. New edition 2006.

Copyright © 2006 The Open University

All rights reserved; no part of this publication may be reproduced, stored in a retrieval system, transmitted or utilised in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without written permission from the publisher or a licence from the Copyright Licensing Agency Ltd. Details of such licences (for reprographic reproduction) may be obtained from the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP.

Open University course materials may also be made available in electronic formats for use by students of the University. All rights, including copyright and related rights and database rights, in electronic course materials and their contents are owned by or licensed to The Open University, or otherwise used by The Open University as permitted by applicable law.

In using electronic course materials and their contents you agree that your use will be solely for the purposes of following an Open University course of study or otherwise as licensed by The Open University or its assigns.

Except as permitted above you undertake not to copy, store in any medium (including electronic storage or use in a website), distribute, transmit or re-transmit, broadcast, modify or show in public such electronic materials in whole or in part without the prior written consent of The Open University or in accordance with the Copyright, Designs and Patents Act 1988.

Edited, designed and typeset by The Open University, using the Open University TEX System.

Printed in the United Kingdom by Thanet Press Ltd, Margate.

ISBN 0 7492 0275 0

3.1

Contents

Guidance notes 4

Chapter D1 5

Section 2 Exploring the problems 5

Chapter D2 14

Section 3 Fitting a normal model 14

3.1 Introducing OUStats 14

3.2 Is a normal model a good fit? 23

3.3 Printing with OUStats 29

Section 4 Are people getting taller? 30

Section 5 Exploring normal distributions 38

Chapter D3 41

Section 3 Confidence intervals on the computer 41

3.1 Interpreting a confidence interval 41

3.2 Calculating a confidence interval 44

Chapter D4 47

Section 2 Exploring the data 47

Section 4 Testing for a difference 50

Section 6 Fitting a line to data 53

Appendix: Entering and editing data 57

Solutions to Activities 59

Index for OUStats 69

Acknowledgements 70

Guidance notes

This computer book contains those sections of the chapters in Block D which require you to use your computer. Each of those chapters contains instructions as to when you should first refer to particular material in this computer book, so you are advised not to work on the activities here until you have reached the appropriate points in the chapters.

For advice on how each computer session fits into suggested study patterns, refer to the Study guides in the relevant chapters. The statistical software for Block D will have been installed when you installed the MST121 software in preparation for Block A. This block does not draw on Mathcad.

About the software

The software for Block D has two components: *Simulations* and *OUStats*. You will be using *Simulations* in Chapters D1 and D3. The main component of the software is *OUStats*. This is a data analysis package, which is introduced in Chapter D2 and used in each of the remaining chapters of the block.

In this block, and in particular when using *OUStats*, it is assumed that you are familiar with the following topics.

Frequency diagrams
Median and quartiles
Boxplots
Mean
Standard deviation
Scatterplots

Frequency diagrams are needed from the start; they are used in Section 1 of Chapter D1. The mean is first mentioned in Section 4 of Chapter D1; the standard deviation and scatterplots are used in Chapter D2 (in Sections 2 and 4, respectively). The median, quartiles and boxplots are used in Chapter D4 after being reviewed briefly in Section 1 of that chapter.

Chapter D1, Section 2
Exploring the problems

The first activity takes you through some of the basic features of the probability simulations which are provided as part of the statistics software.

Activity 2.1 Probability simulations

(a) To locate the *Simulations* package, click on the **Start** menu, move the mouse pointer to **Programs**, then click on **MST121 Simulations**. After a pause, you should see a 'Home' window featuring the following options.

Experiments	Waiting for a success
Settling down	Collecting a complete set
Heads	Confidence intervals

Each of these is accessible via a tab near the top of the window or via a panel towards the right-hand side.

In this section, you will be using all of these simulations except the last; Confidence intervals will be used in Chapter D3.

When you wish to leave the package, click on the **File** menu and then **Exit...** (or simply click on the 'Close' button at the top right-hand corner of the window).

Alternatively, you can access *Simulations* directly by double-clicking on the corresponding icon on your desktop.

(b) Click on **Experiments** (either the tab or the panel) to open this simulation, and you should see the window shown in Figure 2.1.

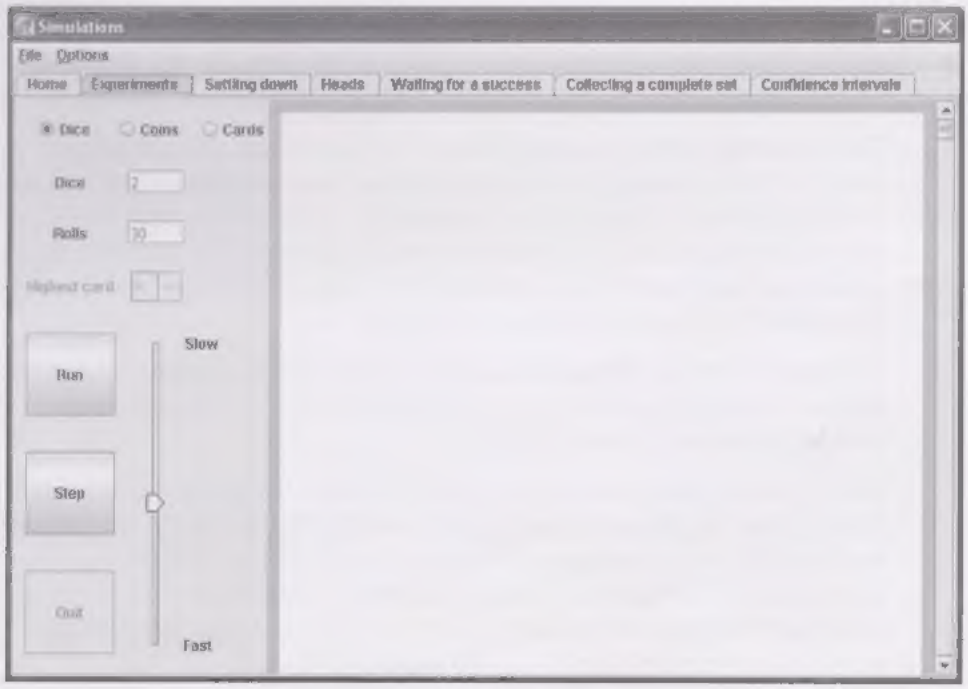


Figure 2.1 The opening window for the **Experiments** simulation

Near the top left of the window are three buttons, labelled **Dice**, **Coins** and **Cards**. The default option is **Dice**. A different option can be selected either by clicking on its name (**Coins** or **Cards**) or by clicking on its button.

Select **Coins**.

Now enter the following settings in the boxes near the top left of the window.

Coins [1]

Tosses [30]

The procedure for doing this is explained below, in case you are not sure how to do it.

You will notice that when the option **Coins** is selected, you are provided with default values for the number of coins and the number of tosses; these values are 2 and 30, respectively. Clicking once in a box positions the cursor in the box. The number in the box can then be edited using the cursor keys, the `[Delete]` key and the `[Backspace]` key. Try typing in a different value for the number of coins now.

You can move the cursor to another box by clicking in the box. Alternatively, if you press the `[Tab]` key, the cursor will move to the next box in the window (in this case, the box for the number of tosses). This number can then be edited as described above. Try doing this now. Then change the settings to 1 coin and 30 tosses.

If you press the `[Tab]` key repeatedly, the cursor selects each of the boxes or buttons within the window in turn. Try this now. You can make use of this feature to run the simulation, or to quit, without using the mouse. For example, if the option marked **Run** is selected in this way, then confirming it by pressing the `[Space]` bar will run the simulation. However, it is usually simpler to run the simulation by clicking on **Run**. Do this now.

Follow the outcomes of 30 tosses of a coin as they appear on the screen. Once the simulation is completed, you can use the scroll-bar (on the right of the window) to scroll back through the outcomes.

(c) Two features of the simulation yet to be mentioned are the speed slider, which is located to the right of the **Run** and **Quit** buttons, and the **Step/Pause** button. By dragging the speed slider with the mouse, the speed of the simulation can be altered. Try running the simulation once more and, while it is running, adjust the speed using the slider. First slow it down, then speed it up.

Dragging a screen object involves placing the mouse pointer on the object and then moving the mouse while holding down the mouse button.

Situated below the **Run** button is a button labelled **Step**. Each time **Step** is clicked on, one toss of a coin is simulated. Click on **Step** several times now to see this effect.

Next, click on **Run** and set the speed to Slow using the speed slider. Notice that the **Step** button is now labelled **Pause**. Click on **Pause**: the simulation is interrupted. You can continue the simulation either by clicking on **Run** or by clicking on **Step**. The effect of clicking on the **Quit** button is to terminate the currently running (or paused) simulation. Spend a few moments exploring these facilities.

(d) Now explore the options **Dice** and **Cards**. When you have finished using this simulation (or indeed any of the simulations), you can return to the 'Home' window by clicking on the **Home** tab near the top left of the window. (Alternatively, you can switch directly to one of the other simulations, by clicking on its tab.) When you have finished exploring the options within the **Experiments** simulation, click on **Home**.

If nothing happens when you click on **Dice** or **Cards**, then check that your previous simulation has finished and is not still running slowly or 'paused'.

You have now explored the first computer simulation. Before going on to use the second one, it is worth reflecting on the art of designing a 'good' computer simulation. On the one hand, a software designer will wish to exploit the power of the computer in order to reduce repetitive routine calculations and tasks, and present only a distilled summary of the reality being simulated. On the other hand, if what you see on the screen differs too much from this reality (the repeated tossing of coins, or whatever), the result can seem abstract and confusing.

The simulation which you have just run is intended as a sort of halfway house between reality and abstraction. It has the advantage of remaining close to the real-world activity of tossing coins, but it does not exploit the computer fully. The remaining simulations, which are explored in the following activities, exploit the computer more effectively, although the particular situation being simulated may not always be obvious from the screen. As you work through the activities which follow, make sure that you understand what each simulation represents.

Activity 2.2 invites you to use the **Settling down** simulation to explore the 'settling down' phenomenon observed in Activity 1.1 of Chapter D1.

Activity 2.2 Settling down

(a) Open the **Settling down** simulation and you will see the window shown in Figure 2.2.

The command in the **Options** menu allows you to choose **Thick lines...** for this simulation.

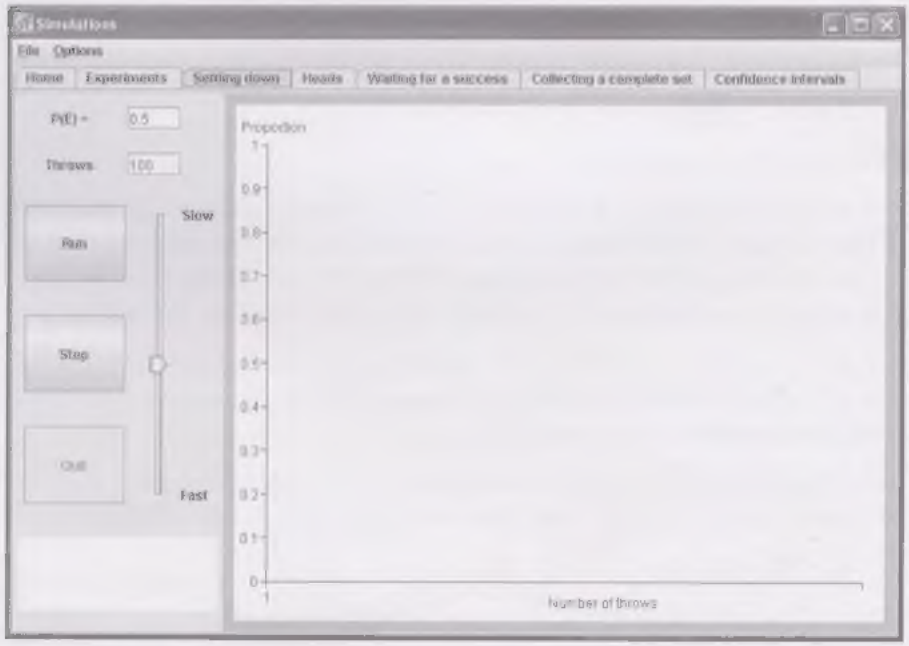


Figure 2.2 The opening window for the **Settling down** simulation

The first click on **Step** generates *two* simulated throws.

(b) The default number of throws is 100. Reset the number of throws to 30. Now click on **Step** and observe the outcome in the bottom left part of the screen. Notice how the result is displayed on the graph. Click on **Step** several times, checking as you do this how the graph corresponds to the data at each step. The simulation is 'doing' what you did with a real coin in Activity 1.1 of Chapter D1.

If you now click on **Run** to complete the simulation, the graph will be completed. Note that at any time you can alter the speed at which the simulation runs (by dragging the speed slider). You can also interrupt the simulation and step through it at any point during its run (by clicking on **Pause** and then on **Step**). Thereafter, you can return to **Run** at any time. Clicking on **Quit** will terminate the simulation.

(c) Now increase the number of throws to 100, and run the simulation once more. Run the simulation for 500 throws and then for 1000 throws. Is the settling down effect apparent in your simulations?

Comment

We cannot predict precisely the results of your simulations. However, it is likely that the settling down effect did become more marked as the number of throws increased.

Activity 2.3 The Brains Trust

Dr Joad defined the law of averages as follows:

> if you spin a coin a hundred times, it will come down heads fifty times, and tails fifty times.

The **Heads** simulation can be used to simulate tossing one or more coins up to 1000 times. The maximum number of coins allowed by the simulation is 20. Use this simulation to investigate the number of heads obtained when one coin is tossed 100 times. Then read the comment below.

Comment

Throughout this computer book, 'I' refers to a particular member of the course team who carried out these activities. The results reported have no special status, but can be used to provide a counterpoint to what you found, as well as enabling discussion of specific points that arise from the particular data obtained.

Using the **Heads** simulation, I entered the following settings.

Coins 1

Tosses 100

Then I ran the simulation.

There are two buttons, **Frequencies** and **Proportions**, near the top left of the window. When **Frequencies** is selected, the frequency is displayed at the top of each bar on the graph. When **Proportions** is selected, proportions are displayed. (To select an option, click on its button.)

For my simulation, the frequency displayed on the '1 head' bar was 44. When I selected **Proportions**, the proportion of tosses which resulted in a head was displayed: this was 0.44.

I ran the simulation a further nine times, and each time noted the frequency on the '1 head' bar. The ten frequencies were as follows.

44 49 51 48 55 46 53 46 45 52

None of my ten simulations produced exactly 50 heads, which appears to knock Dr Joad's definition firmly on the head! However, the average number of heads obtained in these ten runs is 48.9, which is quite close to 50. So the average proportion of tosses which resulted in a head was close to $\frac{1}{2}$.

Perhaps Dr Joad's definition of the law of averages could be reworded as follows:

> if you spin a coin a large number of times, the proportion of spins that result in a head will be approximately $\frac{1}{2}$.

Activity 2.4 D'Alembert's heads

D'Alembert argued that, in two tosses of a coin, there are three possible outcomes – heads on the first toss, heads on the second toss, and heads on neither toss. By his reasoning, since two of these three give at least one head, the probability that the coin lands heads at least once is $\frac{2}{3}$.

(a) Use the **Heads** simulation to investigate d'Alembert's conclusion. This time two coins are being tossed, so set the number of coins to 2. You may need to run the simulation several times, with various different numbers of tosses, in order to reach a conclusion. Record your results in a table like the one below, and write down your conclusions.

Remember that, for our purposes, tossing two coins is equivalent to tossing one coin twice.

Run	Number of tosses	Tosses which gave at least 1 head	
		Number	Proportion
1	100		
2			
3			

(b) Do you think that d'Alembert's conjectured probability of $\frac{2}{3}$ is correct? If not, having carried out some simulations, what do you think the correct value of the probability is? Do the results of your simulations agree with the ideas you jotted down in Subsection 1.3 of Chapter D1?

Comment

We shall return to this problem in Section 3 of the main text.

Activity 2.5 Waiting for a six

In some board games, players can join in only when they roll a six with a die. In Subsection 1.3, you were invited to write down your ideas concerning several questions about the length of time (measured as the number of rolls of a die) that a player has to wait to join in a game. The **Waiting for a success** simulation can be used to investigate these questions.

(a) Open the **Waiting for a success** simulation. Each time a die is rolled, the probability of obtaining a six is $\frac{1}{6}$. If we regard obtaining a six as a success, then $P(\text{success}) = \frac{1}{6}$. The number of times the die has to be rolled to obtain a six (a success) is the wait. On the screen, enter the following values for the settings: $P(\text{success}) = 1/6$ and 1 wait.

Run the simulation several times, and try to get a sense of what lengths of wait, typically, tend to occur.

(b) Now set the number of waits to 50, and step through the first few waits to ensure that you understand what is going on. Then click on **Run** to complete the simulation. You should obtain output similar to that shown in Figure 2.3.

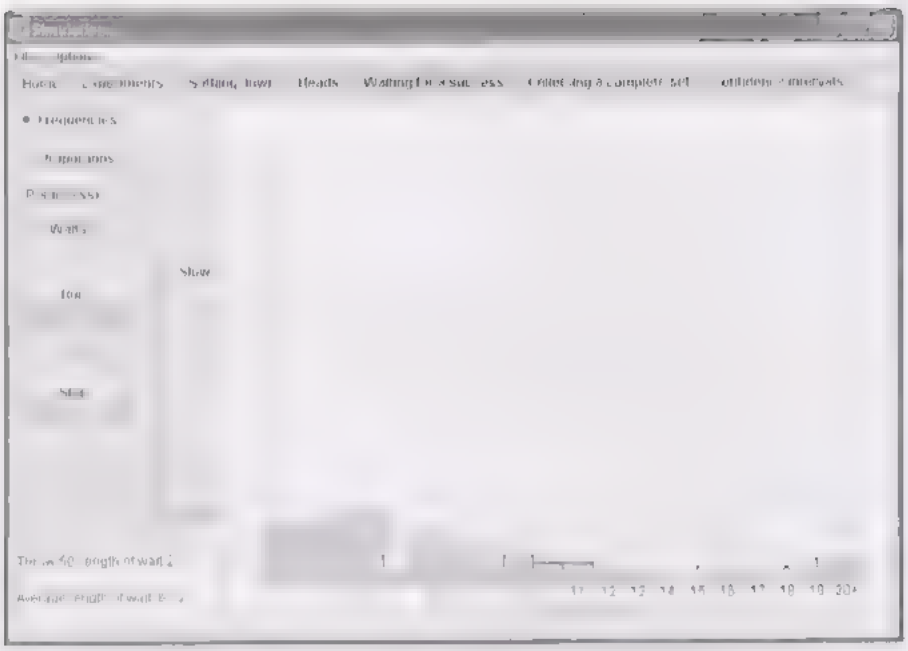


Figure 2.3 The results of a run of the **Waiting for a success** simulation

Notice that if, as in the simulation depicted in Figure 2.3, you obtained some waits that were longer than 19, then these are registered on the '20+' bar at the right-hand end of the horizontal axis.

Notice also the extra information that appears at the end of a simulation in the box in the bottom left part of the window. This is the average length of the waits in the simulation just run.

(c) Run the simulation several times. Can you tell from your simulation when you are most likely to achieve a six? That is, what number of rolls is most likely to be needed to obtain a six?

Comment

Overall, you may have found that your results were very variable and it was difficult to draw any firm conclusions based on just 50 waits. A greater number of waits is clearly necessary. You are asked to try this in the next activity.

Activity 2.6 Still waiting for a six

In this activity, you are invited to continue your investigation from Activity 2.5 by running the simulation a number of times for a larger number of waits. As you do so, focus on the following questions.

- On average, how many times will a player have to roll a die in order to obtain a six?
- What is the most likely number of rolls needed to obtain a six?

You may find it helpful to record your results in a table like the one below.

Number of waits	Average wait	Most likely number of rolls
300		
300		
300		

(a) Run the simulation several times using 300 waits. On each occasion, note the average length of the waits and the number of rolls (that is, the length of wait) which occurred most frequently. What do you notice about the frequencies of the different wait lengths? How would you describe the general shape of the frequency diagram?

(b) Use your results to form hypotheses about the answers to the two questions above. Experiment with different numbers of waits to help you do this.

(c) How do your hypotheses compare with your intuitions? Were you surprised by any of the results that you obtained?

Comment

The problem *Waiting for a six* is investigated further in Section 4 of the main text.

Activity 2.7 How long is an average wait?

The **Waiting for a success** simulation can be used to investigate the waiting time for other events; for example, the number of tosses of a coin needed to obtain a head, or the number of children a couple might need to have to produce a girl. If we assume that $P(\text{head}) = \frac{1}{2}$ and $P(\text{girl}) = \frac{1}{2}$, then in both examples we can use the simulation with $P(\text{success}) = \frac{1}{2}$. Other events would require different values of $P(\text{success})$.

(a) Use the simulation to explore the average wait for various values of $P(\text{success})$. Note down your results, and use them to predict a value for the average wait for each value of $P(\text{success})$ that you choose. You may find it helpful to record your results in a table like the following one.

$P(\text{success})$	Number of waits	Average wait: observed values	Average wait: prediction

(b) Can you spot any pattern in your results? If $P(\text{success}) = p$, what would your conjecture be for the average wait?

Comment

We shall return to this problem in Section 4 of the main text.

The final two computer activities in this section use the **Collecting a complete set** simulation. It has been designed to allow you to investigate the problem *Collecting a complete set of musicians.*

Activity 2.8 Collecting a complete set of musicians

One out of eight different toy musicians is given away in each packet of a popular breakfast cereal. In Subsection 1.3 of Chapter D1, you were asked to guess the number of packets of cereal that you might expect to have to buy in order to collect a complete set of eight musicians. In this activity, you are invited to investigate this problem using the **Collecting a complete set** simulation.

(a) After opening the window for this simulation, change the number of objects in a set to 8 (and leave the number of collections at 1). To make sure that you understand what this simulation does, step through the simulation until you obtain a complete set. At each step, each object has an equal chance of being selected. Objects are selected until at least one of each different type has been chosen. The number of the last object selected is highlighted on the horizontal axis and recorded in the box at the bottom left of the window. The number of packets needed to complete the collection is eventually displayed in this box also.

If you run the simulation for a number of collections greater than 1 then, when the simulation finishes, all of the results are recorded in the box at the bottom left of the window. The number of packets needed to complete each collection is displayed after 'Packets:'. If the results are not all visible, then they can be viewed by scrolling through them. If you click on a line in this box, representing a particular collection, then the corresponding diagram is displayed.

(b) Now run the simulation several times, each time noting the number of packets required to obtain a complete set. Run the simulation so as to obtain at least 10 collections, and write down the number of packets that were required for each collection. (You will need to refer to your results again in Section 5 of the main text.)

Number of packets required to collect a complete set

(c) Did you find your results surprising? Are they consistent with the ideas you noted in Subsection 1.3? Do you wish to revise the conjecture you made in Subsection 1.3 for the average number of packets required to collect a complete set? If so, then make a note of your revised prediction.

Comment

What you might have found striking was the great variability in the number of packets needed to collect a complete set of eight musicians. In Section 5 of the main text, we return to the problem of finding the average number of packets required.

Activity 2.9 Collecting a complete set

The **Collecting a complete set** simulation can be used to explore the numbers of packets needed to collect complete sets of sizes other than 8. Consider the following situations.

(a) Take a well-shuffled pack of cards, and cut the pack at random, noting the card revealed. Repeat this procedure until you have selected at least one card from each suit. How many times would you expect to have to cut the pack?

(b) You toss a coin repeatedly, noting the result (head or tail) each time. You stop at the point when you first have at least one occurrence of each outcome. How many tosses would you expect to have to make?

Explore each of these questions using the **Collecting a complete set** simulation.

Chapter D2, Section 3
Fitting a normal model

3.1 Introducing OUStats

In this subsection, you will be guided through some of the facilities of *OUStats*, the data analysis part of the statistics software.

Nearly all the data sets that you will be exploring and analysing in this chapter (and in the rest of Block D) are large, and simply typing the data into *OUStats* would occupy a lot of time. In consequence, all the data sets are provided as data files. You will not be asked to enter data sets yourself, either in this block or in assignments, so you will not find any instructions here for entering or editing data. However, instructions have been included in an Appendix in case you wish to be able to use *OUStats* to analyse your own data (or in case you are interested to know how this is done). Whether or not you study the Appendix is entirely up to you.

Start up *OUStats* now, as follows.

◇ Click on the **Start** menu, move the mouse pointer to **Programs**, then click on **MST121 OUStats**. The following window should appear.

Alternatively, you can access *OUStats* directly by double-clicking on the **MST121 OUStats** icon on your desktop.

Instructions for 'clicking' here and below refer to use of the left-hand mouse button, unless stated otherwise.

The underlined letters in menu titles indicate the keystroke with `[Alt]` that can be used (instead of a mouse click) to open the menu. Thus `[Alt]F` will open the **File** menu.

Note that you can exit from *OUStats* at any time, simply by clicking on **File** and choosing **Exit** (by clicking on it). Alternatively, click on the 'Close' button at the top right-hand corner of the window.

Figure 3.1 The opening window in *OUStats*

This is a blank data window, with title 'untitled1.ous'. It is made up of cells, with rows numbered $1, 2, 3, \ldots$ and columns labelled V1, V2, V3, ...

Across the top of the window are seven menu titles: **File**, **Edit**, **Stats**, **Plot**, **Options**, **Window** and **Help**. Four of these (**File**, **Edit**, **Window**, **Help**) should look familiar from other *Windows*-based packages.

◇ Click on **File**.

As well as the menu commands for handling files

◇ Click on **Edit**.

The commands in this menu allow you to edit the data window, rename columns, and so on.

◇ Click on **Window**.

The first two commands here, **Cascade** and **Tile**, provide different rearrangements of the windows that you create.

◇ Click on **Help**

The first command gives access to the on-screen Help facility.

The other menus, **Stats**, **Plot**, **Options**, are specific to this software package.

◇ Click on **Stats**.

The commands here provide for a range of calculations to be carried out on data. To use most of them, you need first to open or create a data file. You will be asked to open such a file shortly, and the use of some of these commands will then be demonstrated. Other commands will be used in later chapters: **Confidence interval...** in Chapter D3, **Two sample z-test...** and **Regression...** in Chapter D4.

◇ Click on **Plot**.

This menu contains the commands for obtaining diagrams to represent

Frequency diagram...

Scatterplot... will be used later in this chapter, and **Boxplot...** in Chapter D4.

◇ Click on **Options**.

The first two commands here allow you to change some of the settings of the package, and in particular the number of significant figures to which output values are displayed.

This completes a first look at the options available from the seven menus. You are next asked to open a data file.

◇ Click on **File**. Then choose **Open...** (by clicking on it).

All of the file names shown have the extension '.OUS', indicating that they are *OUStats* data files.

◇ Scroll through the list until you find the name GEYSER.OUS. Click on this name (adding it to the **File Name** box), then click on **Open**

You should now have a data window that contains a single column of data, headed 'Intervals'. When a data file is open, you can obtain information about the data using **Notes...** in the **File** menu.

◇ Click on **File**, and select **Notes...**. Read what this text says about the data in GEYSER.OUS. Then remove the 'Notes' window by clicking on the 'Close' button at its top right-hand corner.

You will next see how to obtain summary statistics and a frequency diagram for this data set. This will show how intervals between eruptions of the geyser vary.

◇ Click on **Stats**, then select **Summary stats...**. In the window that appears, select 'Intervals' (by clicking on it) and click on **Select**.

Alternatively, use the keystroke combination [Alt]F

Bar chart... and **Pie chart...** are not used in the course, but you may wish to explore their use at another time.

After a possible pause, summary statistics for the data set appear lower in the window. These include the mean, standard deviation, minimum value, and so on. The last item is the sample size, which is 299. Also shown are the median and quartiles; these will be reviewed in Chapter D4. Note (for reference shortly) that the minimum and maximum values are, respectively, 43 and 108 (in minutes). Next we obtain a frequency diagram.

◇ Click on **Plot**, then select **Frequency diagram...**. From the left-hand drop-down menu, select 'Intervals' (which is the only item in this case). Then click on **Go**.

The frequency diagram that appears in the window uses a first interval starting value and interval width that have been generated automatically by the software. These values can also be specified by the user. We noted earlier that the data had minimum value 43 and maximum value 108. These values suggest that a first interval starting value of 40 and interval width of 10 might be suitable.

The second box can be reached either by clicking on it or by using the `[Tab]` key.

◇ Click on the **First interval** box. Edit the contents so that 'auto' is replaced by 40. Then edit the **Width** box so that 'auto' is replaced by 10. Finally, click on **Go**.

The frequency diagram is redrawn using these values. You now have three open windows: the original data window, a 'Summary statistics' window and a 'Frequency diagram' window. It is possible to see all of these together, in an orderly format, using the **Tile** facility.

◇ Click on **Window**, then select **Tile**.

The three windows now each occupy a quarter of the overall *OUStats* window, with a blank space in the remaining quarter.

This concludes our look at output arising from the data in GEYSER.OUS. In order to close the file, it suffices to open another one, which will be used to demonstrate further features of *OUStats*.

◇ Click on **File**. Then choose **Open...**. Locate the file name HEIGHTS.OUS and click on it, then click on **Open**.

The file HEIGHTS.OUS contains frequency data.

◇ Click on **File**, and select **Notes...**. Read what this text says about the data in HEIGHTS.OUS.

First we obtain summary statistics for these data.

◇ Click on **Stats**, then select **Summary stats...**

The list of variables here contains three items, 'Height', 'Frequency' and 'Height | Frequency'. The first two just correspond to the values held in the columns of the same names, but 'Height | Frequency' takes account of the linkage between the two columns (that is, that a height of 62 inches occurs 3 times, 63 inches occurs 20 times, and so on). Hence it is the last choice which is the correct one here.

◇ Choose 'Height | Frequency', then click on **Select**.

The sample size is 1000 (Cambridge men). The values range from a minimum of 62 inches to a maximum of 77 inches. (Note these two values for reference shortly.) Now we turn to a frequency diagram.

◇ Click on **Plot**, then select **Frequency diagram...**. From the left-hand drop-down menu, select 'Height | Frequency'. Set **First interval** to 60 (slightly less than the minimum of 62 noted above) and set **Width** to 2. Finally, click on **Go**.

The shape of the frequency diagram (which is the same shape as that of the corresponding histogram) indicates that a normal model might be suitable for the heights of Cambridge men. So next we fit a normal curve to the data.

You saw this also in Section 1.

◇ In the second drop-down menu from the left within the 'Frequency diagram' window, select 'Fit normal curve'.

The display changes from a frequency diagram to a histogram. Also new boxes appear, indicating the mean (68.872) and standard deviation (2.56792) of the data set. If the **Fit normal curve** button is clicked on, then a normal curve with the indicated mean and standard deviation is added to the histogram. However, it is also possible to edit the values in the **Mean:** and **Std dev:** boxes to more appropriate accuracy (three significant figures, say).

◇ Change the value in the **Mean:** box to 68.9. Then change the value in the **Std dev:** box to 2.57. Finally, click on the **Fit normal curve** button.

Note these numbers for use shortly.

The chosen normal curve is now superimposed on the histogram. It does indeed seem to fit the data well. The curve is a model for the heights of all Cambridge men in 1902. It can be used to estimate, for example, the proportion of Cambridge men in that year who were between 69.5 and 70.5 inches in height. To find this proportion, we seek the area under the curve between 69.5 and 70.5. This can be found using **Normal distribution...** from the **Stats** menu.

◇ Click on **Stats**, then select **Normal distribution...**. Within the window that appears, change the value in the **Mean** box to 68.9, and (after pressing [Tab] to change boxes) change the value in the **Standard deviation** box to 2.57 (both as noted above). Then click on **Update** to update the graph below.

The normal curve shown corresponds to the mean and standard deviation just entered. Above the curve (and below the **Update** button) are four boxes and two buttons. The first two boxes, **A** and **B**, show their preset values (which are both 0). The two lower boxes show the corresponding values of **Area to left of A** and **Area between A and B** (which are currently both displayed as 0). The window can now be used in either of two ways:

(a) to input values for A, B and find the corresponding areas;

(b) to input values for the areas and find the corresponding values for A and B.

The area to the left of A is actually non-zero but extremely small, since A = 0 is many standard deviations below the mean.

To demonstrate the first of these, we find the area under the curve between 69.5 and 70.5.

◇ Click on the **A** box (or use [Tab] repeatedly to reach this by navigation around the boxes and buttons). Change the value in this box to 69.5. Press [Tab] to move to the **B** box, and change the value here to 70.5. Now click on **Use A & B to calc areas**.

The area between A = 69.5 and B = 70.5 is now shown shaded on the graph. The value of this area is displayed in the **Area between A and B** box (and also on the graph). The value of the area under the curve to the left of A is also shown both in a box and towards the left on the graph.

Put another way, if a man had been chosen at random from this population, then the probability that his height would have been between 69 5 and 70 5 inches is 0 141.

The value of the shaded area is 0.141 (to three significant figures). This means that, according to the model, 14.1% of Cambridge men in 1902 were between 69.5 and 70 5 inches tall. Another conclusion here, with reference to the **Area to left of A** box, is that, according to the model, 59.2% of Cambridge men in 1902 were less than 69.5 inches tall.

We now use the window in the second way described above, by finding a value of A that depends on a given value of the area to the left of A. Suppose that we seek the height such that 95% of men in this population were shorter than this height.

◇ Click on the **Area to left of A** box and enter 0.95 into it. Then click on **Use areas to calc A & B**

This is as it should be, since the total area beneath the normal curve is 1.

An error message appears, indicating that the sum of the two areas (that to the left of A, and that between A and B) must be less than 1.

◇ Click on **OK**, and enter in the **Area between A and B** box any non-negative number less than 0.05 (0 suffices). Then click again on **Use areas to calc A & B**

Alternatively, areas can be selected by clicking and dragging with the mouse on the graph. See later for further details

The value obtained in the **A** box is 73.1273. Hence, according to the model, 95% of Cambridge men in 1902 were less than about 73.1 inches tall.

That concludes this introductory tour of some *OUStats* facilities. You may wish to spend some time on your own at this point to explore other features of the package. For reference purposes, the main facilities of *OUStats* introduced so far are summarised below.

(1) The menus

The **File** and **Edit** menus contain commands for handling and editing files.

The **Stats** menu contains commands for calculations.

The **Plot** menu contains commands for obtaining diagrams.

The **Options** menu allows you to change some of the settings of *OUStats*. These include the number of significant figures shown in displayed results and the colours to be used in displayed graphs. You can also alter a 'graphics smoothing' setting in order to improve the appearance of printouts obtained from graph windows. Additionally, the Calculator resident in *Windows* is accessible via this menu.

The **Window** menu contains commands for arranging windows on the screen. (These are standard *Windows* operations.)

The **Help** menu provides access to on-screen help.

(2) Data files

To open a data file:

◇ click on **File** and choose **Open...** (by clicking on it) a dialogue box appears;

◇ click on the file name, then click on **Open.**

Information about the data in the file currently open can be obtained by choosing **Notes...** in the **File** menu.

(3) Summary statistics

To obtain summary statistics for a variable in an open data file:

- ◇ click on **Stats**, then choose **Summary stats...** (by clicking on it);
- ◇ select the variable name(s) from the scrolling list in the new window that appears (by clicking on it/them), then click on **Select**.

To select more than one variable at a time, hold down the [Ctrl] key while making your choice.

The summary statistics are then displayed in the window.

(4) Frequency diagrams

To obtain a frequency diagram:

- ◇ click on **Plot**, then choose **Frequency diagram...** (by clicking on it);
- ◇ select the variable name from the left-hand drop-down menu in the window that appears;
- ◇ enter the first interval starting value in the **First interval** box and the interval width in the **Width** box, should you not want the software to choose these values automatically, then click on **Go**.

Each such diagram produced is displayed in a separate window, titled 'Frequency diagram:' followed by the variable name.

Histograms are produced in a similar way. Choose the second drop-down menu from the left within the 'Frequency diagram' window, and select 'Histogram' in place of 'Frequency'.

(5) Fitting a normal curve to data

To fit a normal curve:

- ◇ obtain a frequency diagram for the data, as above;
- ◇ in the second drop-down menu from the left, select 'Fit normal curve' (the display then changes from a frequency diagram to a histogram);
- ◇ enter the mean and standard deviation of the required normal curve in the appropriate boxes (or accept the values provided), then click on the **Fit normal curve** button.

The default values are the sample mean and sample standard deviation.

The normal curve with the specified mean and standard deviation is then superimposed on the histogram.

(6) Finding areas under a normal curve

To find an area under a normal curve, you must first:

- ◇ click on **Stats**, then choose **Normal distribution...** (by clicking on it), whereupon a window similar to that in Figure 3.2 overleaf will appear;
- ◇ enter the mean and standard deviation of the required normal distribution in the appropriate boxes, and click on **Update**.

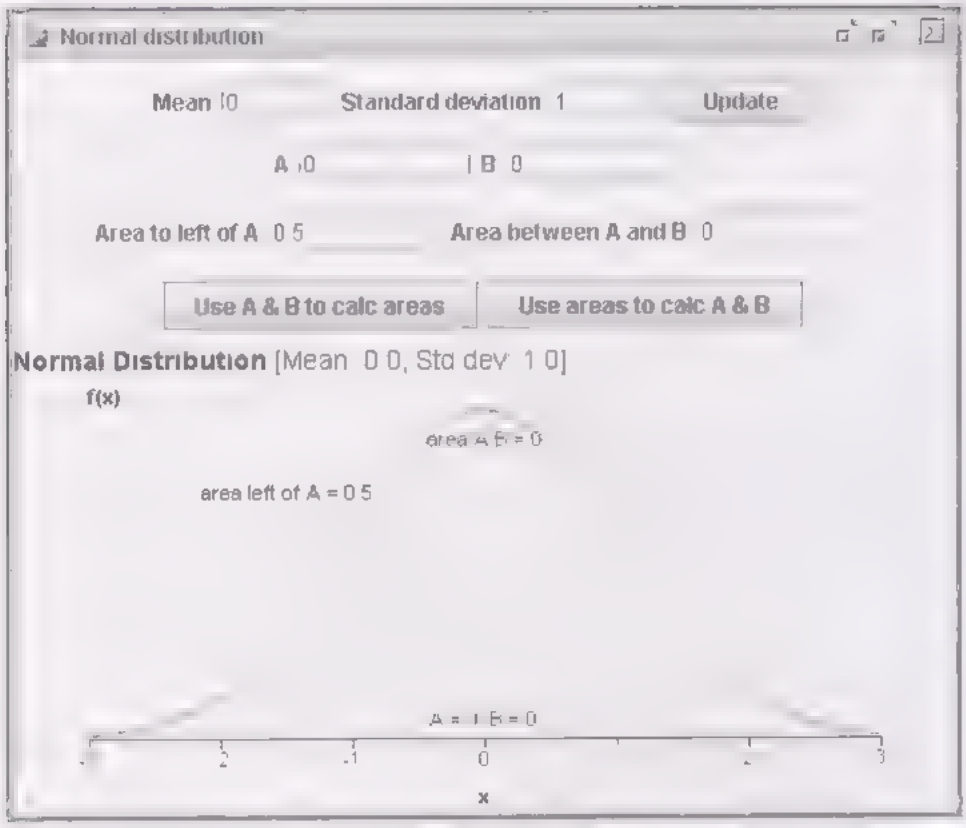


Figure 3.2 The opening window for **Normal distribution...**

To find the area to the left of a numerical value, a (say):

- ◇ *either* edit the value in the **A** box to read a and the value in the **B** box to be greater than a, then click on the **Use A & B to calc areas** button;
- ◇ *or* position the mouse pointer close to the horizontal axis, hold the mouse button down and drag the mouse until the value in the **A** box is as close as possible to a, then let go of the mouse button.

You may position the mouse pointer anywhere in the window, outside of boxes and buttons, but you will probably find it more useful to position it close to the horizontal axis.

The area to the left of a will appear in the **Area to left of A** box (and also above the graph).

To find the area between two values, a and b (say):

- ◇ enter the value a in the **A** box and the value b in the **B** box by either of the methods just described.

Dragging the mouse from left to right increases the value of B (from a starting value of A) when the mouse is released. Dragging from right to left decreases the value of A.

If you choose to press and drag the mouse then the area under the curve between a and b is shaded as soon as you let go of the mouse button and the area is given in the **Area between A and B** box (and also above the graph). If you edit the values, then the area is shaded and its value is given as soon as you click on the **Use A & B to calc areas** button.

To find the value a such that the area to the left of a is equal to a numerical value p (say):

- ◇ edit the **Area to left of A** box so that it contains the number p;
- ◇ click on the **Use areas to calc A & B** button.

The value of p must be less than 1. Also, the value in the **Area between A and B** box should be less than $1 - p$.

The value a will be displayed in the **A** box.

(7) Saving files and copying from windows

If you make any changes to one of the data files supplied with *OUStats*, and wish to save the amended data window, then you must do so using **Save As...** from the **File** menu, and you must save it using a different name: all the data files and notes files supplied are protected so that you cannot accidentally change the contents of the original files.

The computer operating system does not distinguish between upper- and lower-case letters in file names: you can use either. For clarity of presentation, we have used upper-case letters in file names in this computer book.

Active output from an *OUStats* window can be copied to other applications in the usual way, using Copy and Paste.

Any window other than the original data window can be closed using the 'Close' button at its top right-hand corner. The data window can be closed only by: (a) opening another data file; (b) creating a new data file; (c) exiting from the application.

When a data file is opened, any previously-saved files are closed down automatically. You will receive a prompt inviting you to save the data window if it has been changed. No windows other than the data window and accompanying Notes can be saved.

(8) Using frequency data

Some data are stored in two columns in a data file, with values in one column and corresponding frequencies in the next. When using such data to obtain summary statistics or a frequency diagram, for instance – note that the two column names appear on a single line in variable lists, with a vertical bar between them: for example, 'Value|Frequency'. However, the individual column labels (in this case 'Value' and 'Frequency') also appear in variable lists as potential choices, so it is important to remember to choose the option with the vertical bar where appropriate.

A comment on frequency diagrams

Before you move on to the next group of computer activities in Subsection 3.2, there is one point concerning frequency diagrams that ought to be mentioned. When using a computer package, it is all too easy to obtain a frequency diagram without giving much thought to whether the diagram you obtain is the 'best' possible. Different choices of starting values and interval widths will, in general, produce different diagrams representing the same data; and not all choices of the starting value and interval width will necessarily be appropriate for the data.

Consider, for instance, the two frequency diagrams for the heights of 1000 Cambridge men that you produced earlier. The first was obtained by allowing *OUStats* to select automatically the starting value of the first interval and the width of the intervals. For the second, we chose these values ourselves: since the heights ranged from 62 to 77, it seemed reasonable to choose 60 as the starting value and 2 as the interval width. But were these good choices?

You may recall that the heights were recorded to the nearest inch. (This information is given in the Notes that accompany HEIGHTS.OUS and was mentioned when the data were introduced in Section 1.) This means that, for instance, the heights of men who were anywhere between 61.5 and 62.5 inches tall were recorded as 62 inches, the heights of men between 62.5 and 63.5 inches tall were recorded as 63 inches, and so on.

For the second frequency diagram, we specified that the first interval should start at 60 and have width 2. So *OUStats* included in this interval all heights recorded as at least 60 inches but less than 62 inches; there were none, since the lowest recorded height was 62 inches. The second interval included all heights recorded as at least 62 inches but less than 64 inches, that is, all those recorded as either 62 inches or 63 inches; there were 23 of these. So there were 23 men between 61.5 and 63.5 inches tall. The heights of these men were represented on the frequency diagram by a bar drawn from 62 to 64, when clearly a bar drawn from 61.5 to 63.5 would have been better.

This sort of discrepancy can be avoided by noting how the data were recorded and choosing intervals appropriately. In this case, since the shortest recorded height was 62 inches and heights between 61.5 and 62.5 inches were recorded as 62, it would be sensible to choose 61.5 as the starting value of the first interval, rather than 60 or 62. Then the first bar would be drawn from 61.5 to 63.5. By the way, the frequency diagram in Figure 1.2 of Chapter D2 can be obtained by using a starting value of 61.5 for the first interval and an interval width of 1.

The important message to be obtained from this example is that when you use a statistics software package, you need to think about what you are doing. Although a package will save you all the work involved in doing calculations and drawing diagrams, it will not think for you. It will usually do whatever you ask it to, whether or not your instructions are sensible or appropriate.

3.2 Is a normal model a good fit?

In Subsection 3.1, a normal curve was chosen to model the variation in the heights of Cambridge men in 1902, but no check was made on whether the model was a 'good' fit. In this subsection, we use *OUStats* to investigate informally, for several samples of data, whether a fitted normal curve is a good model for the variation observed in the data.

The first stage in investigating whether a normal distribution is a suitable model should always be to obtain a frequency diagram for the data, and to inspect its shape. If it is clearly not bell-shaped – for example, if it is skewed or has more than one clear peak, such as for the four frequency diagrams in Figure 1.4 of Chapter D2 – then a normal model can be rejected immediately. But if it looks as though a bell-shaped curve might be a suitable model for the variation in the data (as in Figure 1.5), then the next step is to fit a normal curve with parameters μ, estimated by the sample mean $\bar{x}$, and σ, estimated by the sample standard deviation s. The fit of the curve can then be inspected by eye.

Sometimes (as was the case for the heights of Cambridge men) a histogram for the data and the fitted normal curve are so similar in shape that it is clear that the model is a good one. But more commonly, perhaps because of the jaggedness of the histogram, there is some doubt about the fit. The problem is to decide whether the differences between the histogram and the curve could be the result of chance, and just a feature of the particular sample, or whether they are an indication that the model is not a good fit.

There are formal statistical tests that can be carried out, called goodness-of-fit tests, for deciding whether a chosen model is a good one for the variation in a sample of data; and if you study statistics in the future, then you will almost certainly meet such tests. However, in this course, we adopt a more informal approach, using simulations to generate samples from the chosen normal distribution. This gives an indication of the nature of the variation that occurs by chance, and thus offers a benchmark against which to judge whether the data could reasonably be thought to be a sample from the chosen normal distribution.

In the following activities, you are invited to explore whether a normal model is a good fit for each of a number of data sets. In the first activity, you are asked to decide for each data set, by looking at a frequency diagram, whether a normal curve is even worth considering. In each of the subsequent activities, you are asked to fit a normal model, and then use simulations to investigate the suitability of the model.

Activity 3.1 Is a normal model worth considering?

In this activity, for each data set, you should obtain a frequency diagram for the data and hence decide whether or not a normal distribution might be suitable for modelling the observed variation. For each data set, you should go through the following steps:

- ◇ open the data file;
- ◇ read the information about the data contained in **Notes...**;
- ◇ obtain a frequency diagram for the data, taking particular care over your choice of the first interval starting value and interval width;
- ◇ decide whether or not a normal distribution might be suitable for modelling the observed variation, explaining your decision briefly.

Instructions are given for the first data set only.

(a) The file DIPPER.OUS contains the weights in grams of 198 Irish dipper nestlings at age 6–8 days. Open the data file now. (Choose **Open...** from the **File** menu, click on DIPPER.OUS, and then click on **Open**.) Read the information on these data contained in **Notes...**. (Choose **Notes...** from the **File** menu.)

You may need to scroll through the list of file names to find the file.

The weights are grouped – the individual weights are not given. The groups are listed in the first column of the data window, the midpoints of the groups are in the second column (labelled 'Weight'), and the frequencies are in the third column. To represent these data sensibly on a frequency diagram, you need to start the first interval at 9 and use 2 as the interval width.

Obtain a frequency diagram for the data using these values. (Choose **Frequency diagram...** from the **Plot** menu, click on 'Weight – Frequency' in the left-hand drop-down menu, enter 9 and 2 as the starting value for the first interval and the interval width, respectively, and finally click on **Go**.)

Now consider whether a normal distribution might be suitable for modelling the variation in the data. Questions you might consider include the following. Is the frequency diagram roughly bell-shaped, or is it skewed? Does it have more than one peak?

(b) Repeat the process in part (a) for the data on radial velocities of stars which are contained in the data file RADIAL.OUS.

(c) Repeat the process in part (a) for the data on the lengths of sentences written by H. G. Wells which are contained in the data file AUTHORS.OUS.

(d) Repeat the process in part (a) for the data on the lengths of cuckoo eggs which are contained in the data file CUCKOOS.OUS.

A solution is given on page 59.

In each of the next three activities, you should:

- ◇ fit a normal curve to the data;
- ◇ generate random samples from the fitted normal distribution;
- ◇ compare frequency diagrams for the random samples with a frequency diagram for the data.

Fairly detailed instructions are given in the first activity below. You should follow a similar procedure for the other two.

Activity 3.2 Weights of Irish dipper nestlings

(a) *Fitting a normal curve*

Open the data file DIPPER.OUS, and obtain a frequency diagram for the data using a first interval starting value of 9 and an interval width of 2 (as you did in Activity 3.1).

From the second drop-down menu from the left, select 'Fit normal curve'. The display changes from a frequency diagram to a histogram. New boxes appear, indicating the mean (27.9394) and standard deviation (7.747 04) of the data set, displayed to six significant figures by default.

These statistics are chosen because they are estimates for the population mean and population standard deviation; but it would not be reasonable to suppose or claim six-figure accuracy for these estimates. *As a rough guide, it is reasonable to quote sample statistics to one significant figure more than is given in the data used to calculate them.* The weights of the Irish dipper nestlings are given to two significant figures (roughly), so three-significant-figure accuracy is appropriate for the sample mean and sample standard deviation.

Enter the values 27.9 and 7.75 for the mean and standard deviation of the normal curve. Click on the **Fit normal curve** button, and the normal curve is fitted over the histogram.

It is a good idea to jot down the parameters of the normal curve that you fit, as you will need them again later.

It looks as though a normal curve fits the data quite well, although the frequency diagram is quite jagged. To see whether this is the sort of jaggedness that might be expected to occur by chance, we shall compare this data set with some random samples *of the same size* drawn from the normal distribution that we have fitted to the data. The sample size is important here: because of the 'settling down' effect, we should expect small samples to produce more jagged frequency diagrams than large ones, so we must compare the frequency diagrams of random samples of the same size as the sample of data.

(b) *Obtaining random samples from the fitted normal distribution*

Random samples from a normal distribution are obtained using **Normal samples...** from the **Stats** menu. Click on **Stats** and choose **Normal samples...**; a dialogue box appears. Enter 27.9 for the mean and 7.75 for the standard deviation; these are the parameters of the normal curve that you just fitted. The sample size should be the same as for the data, 198 in this case, so enter 198 in the **Samples** box. To generate three random samples, enter 3 in the **Sets** box. Finally, click on **OK**

The samples are generated and stored in the first three empty columns in the data window: these columns are labelled 'Random1', 'Random2' and 'Random3'

Comparing the random samples and the data

We want to compare the frequency diagram of the data with frequency diagrams for the random samples. We shall use the same intervals for all the samples. Before obtaining the frequency diagrams, we need to decide on the starting value for the first interval; and to do this we need to know the lowest value that appears in any of the samples. This could be found by looking at the data and picking out the lowest value, but this is a tedious exercise for large sample sizes. An alternative is to use **Summary stats...**, since the minimum is one of the statistics displayed.

Choose **Summary stats...** from the **Stats** menu. You can select several variables at the same time by holding down the `[Ctrl]` key. Do this and click on 'Random1', 'Random2' and 'Random3' in turn. Then release the `[Ctrl]` key. Click on **Select**, and summary statistics for all three samples will be displayed lower in the window. Pick out the minimum value for each random sample. (You will need to scroll through the output to see them all.)

The minimum value in one of my samples was 5.725 66. This was the lowest of the three minimum values, so I decided to use 5 as the starting value for the first interval for the frequency diagrams (instead of 9, which was used for the data earlier). Your random samples will be different from mine, so you may need to use a different starting value. Remember that we want intervals 9–11, 11–13, and so on, so the starting value of the first interval must be an odd number.

Now we have all the information that we need. For each of the variables 'Weight', 'Frequency', 'Random1', 'Random2' and 'Random3', proceed as follows. Choose **Frequency diagram...** from the **Plot** menu, and select the variable name from the left-hand drop-down menu in the window. Input your starting value for the first interval and 2 for the interval width, then click on **Go**. Having done this for each of the variables, you should have created four frequency diagrams.

d *Viewing the frequency diagrams*

Use **Tile** from the **Window** menu so that you can see all of the open windows at once. Close down any open window other than the data window and the four frequency diagrams just created. Maximise the size of the overall *OUStats* window. Then use **Tile** again, to see the four frequency diagrams together and at similar size. My four diagrams are shown in Figure 3.3.

The data window cannot be closed.

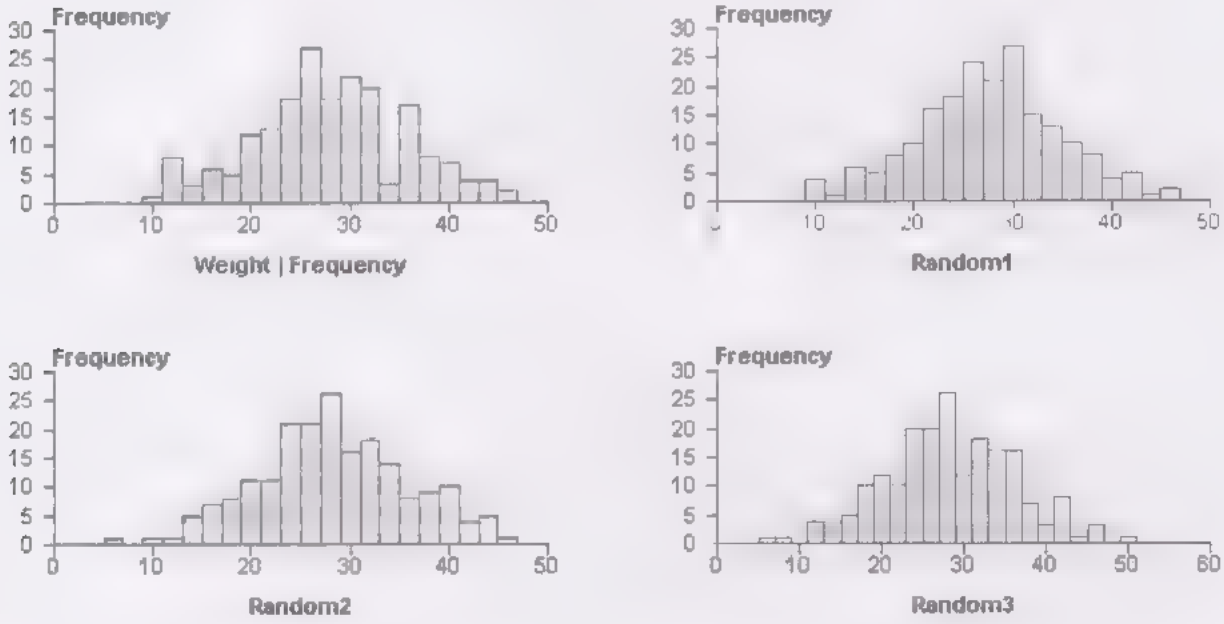


Figure 3.3 Four frequency diagrams from *OUStats*

We are now in a position to compare the variation in the random samples from the normal distribution that we fitted with that in the data, and hence to decide informally whether or not the model is a good fit. As you can see, the variability evident in the frequency diagrams for the random samples is broadly similar to that in the frequency diagram for the data. So it looks as though the normal distribution is indeed a good model for the variation in the weights of Irish dipper nestlings aged 6–8 days.

Your random samples will be different from mine. Are your frequency diagrams similar in shape to those obtained here? Notice that the vertical scales on all these frequency diagrams are the same. However, this may not have been so for your diagrams: the scales depend on the values in the samples being represented, and three of the samples were random samples. When comparing frequency diagrams, you need to note where corresponding scales are different from each other, even within windows of the same overall size. It is possible to change the size of a window, in order to make equal distances on two scales appear to have the same length on screen. Usually this is not necessary for the comparison of frequency diagrams, but since it is a facility which can also be applied usefully in other circumstances, the box below explains how it can be done.

Note that the horizontal scale of the diagram for 'Random3' in Figure 3.3 differs from that of the other three diagrams.

Numerical scales within a resized *OUStats* window continue to feature the same numbers, but the diagram is squashed or stretched along with the window.

Changing the size of a window using the mouse

When the mouse pointer is positioned on the border of a window, a double-headed arrow across the border replaces the mouse pointer on the screen. To make a window wider or narrower, place the mouse pointer on either the left or right edge of the window so that this double arrow is showing. Then press down the mouse button and drag the mouse sideways. When you release the mouse button, the window will be redrawn with the edge in the new position. Similarly, you can make a window taller or shorter by pressing and dragging while the mouse pointer is positioned on either the top or the bottom edge of the window. Any diagram inside the window is resized to fit the new window.

Activity 3.3 Radial velocities

Repeat the steps described in Activity 3.2 to investigate whether a normal model is a good fit for the variation observed in the radial velocities which are contained in the data file RADIAL.OUS.

Some comments are given on page 59.

Activity 3.4 Lengths of cuckoo eggs

Repeat the steps described in Activity 3.2 to investigate whether a normal model is a good fit for the variation observed in the lengths of cuckoo eggs which are contained in the data file CUCKOOS.OUS

Some comments are given on page 60.

Generating random samples: a summary

Random samples from a normal distribution are obtained as follows.

- ◇ Choose **Normal samples...** from the **Stats** menu.
- ◇ Enter the mean and standard deviation of the normal distribution, the sample size (in the **Samples** box) and the number of random samples required (in the **Sets** box), then click on **OK**.

If k samples are generated, then they are stored in the first k available columns in the data window and are named 'Random1', 'Random2', ..., 'Random$k$'. (If further random samples are generated at a later stage, possibly with a different data window but during the same *OUStats* session, then the numbering in 'Random' labels continues from the last number used previously.)

Normal samples and random numbers

The command **Normal samples...**, which is contained in the **Stats** menu of *OUStats*, allows you to generate samples of values chosen randomly from a normal distribution. You may have wondered how this is done.

It may seem paradoxical to use a computer to produce 'random' numbers: we expect any computer program to produce output that is entirely predictable. Nevertheless, computer 'random number generators' are in common use; these generate sequences of 'random' integers. Given an initial value – the *seed* value – the sequence generated is predictable and therefore it is not truly random. Numbers generated in this way are called *pseudo-random* numbers. However, in practice, sequences of pseudo-random numbers are indistinguishable from sequences of random numbers, so they may be regarded as sequences of random numbers and used to simulate random samples in statistical simulations.

Most computer programming languages have a routine that generates pseudo-random integers between zero and the maximum integer N that can be stored by the computer. These integers can be used to generate 'random' values from any distribution – normal, geometric or whatever. The details of how this is done are beyond the scope of this course.

The MST121 statistics software uses your computer's clock to determine the seed value of the underlying random number generator, thus ensuring that each statistical experiment (using *Simulations*) or random sample is different. For example, each time you use the **Normal samples...** command in *OUStats*, the seed value is set using the current date and time, making it extremely unlikely that the samples you obtain are the same as any you have obtained previously.

Note that for lotteries, premium bond draws, etc., random simulations based on a pre-programmed algorithm are not used, as they could be open to discovery. Instead, some physical randomising device is used. One such device is based on the number of electrons moving inside a valve.

3.3 Printing with OUStats

The main steps involved in printing with *OUStats* are set out below.

First, make sure that your printer is connected, is installed under *Windows*, and is switched on. The instructions depend in part on whether you wish to print text windows (containing only text) or graph windows.

To print text windows from *OUStats*:

- activate each window that you want to print by clicking on it (or by choosing its title from the **Window** menu)
- choose **Print...** from the **File** menu, and click on **OK**.

The output from the window will then be printed.

To print graph windows from *OUStats*:

- prior to creating the graph windows, choose **Graph options...** from the **Options** menu and make sure that the **Use graphics smoothing** box is *unchecked*;
- create the windows to be printed, then proceed as above for printing text windows.

The checking of this box gives a clearer image when viewed on screen but a less distinct image when printed on paper.

Alternatively, if a graph window is created before the **Use graphics smoothing** box is unchecked, then after unchecking this box you will need to right-click on the window to obtain the **Refresh** menu, and then to click on this to alter the graphics quality. Following this, proceed as above for printing text windows.

You may like to use the following activity to check that you can print output from *OUStats*.

Activity 3.5 Printing output from OUStats

Open the file DIPPER.OUS.

Calculate summary statistics for the weights of the Irish dipper nestlings; these are displayed in a new window.

Now obtain a frequency diagram for the data; this is displayed in another new window.

Now print the output from each of the two windows that have been produced, following the instructions for printing given above.

Chapter D2, Section 4
Are people getting taller?

In the second half of the 19th century, considerable interest developed in the inheritance of characteristics, both in plants and in animals and humans. In the 1890s, Karl Pearson (1857–1936) determined to obtain data on three physical measurements – height, span of arms and length of left forearm – for a large number of families. Many of the data were collected by college students, some of whom made measurements on as many as twenty families. The data were collated by Dr Alice Lee, a colleague of Pearson's at University College, London; she calculated various statistics and prepared some 78 tables of data. According to Pearson, 'this occupied her spare time for nearly two years'. In 1903, several of these tables were published in an article in the journal *Biometrika*.

K. Pearson and A. Lee, 'On the laws of inheritance in man', *Biometrika* 2 (1903) pages 357–462.

The box opposite is a verbatim extract from the instructions given to those who collected the data.

The instruction sheet also contained diagrams illustrating the second and third measurements described. The data cards on which the measurements were recorded emphasised that 'both father and mother are absolutely necessary and should not be over 65 years of age' and that neither parent should be a step-parent. All measurements were recorded to the nearest quarter of an inch, although the heights were rounded to the nearest inch before tabulation. A great deal of thought went into the instructions and the design of the data cards. For example, experiments were carried out into the effect of wearing boots on measured heights, and as a result it was decided to subtract an inch from the recorded height of each boot-wearer. As well as noting the wearing of boots, collectors were also asked to put L, A or C against all the measurements if a person being measured had ever broken a leg, arm or collar-bone.

All those measured were between 18 and 65 years old. Pearson explained his choice of this restriction on age in the article. He observed that full growth may not be reached until age 25 or thereabouts. However, he realised that insisting that all sons and daughters should be over 25 years old might make collecting the data much more problematic, not least because it might be difficult to interest college students in the project as most of them were aged between 19 and 22.

There was also the fact that fewer families with all the sons and daughters over 25 years old had both parents surviving. So, since growth between 18 and 25 is very small, he fixed on 18 years as the lower age limit. He also observed that, because of the phenomenon of shrinkage with age, it would have been better to take a lower maximum age than 65 years for parents, but this too would have limited the number of available families.

Altogether, over a thousand families were measured. The heights of 1078 father–son pairs and 1375 mother–daughter pairs were included in the results.

FAMILY MEASUREMENTS

Professor KARL PEARSON, of University College, London, would esteem it a great favour if any persons in a position to do so, would assist him by making one set (or if possible several sets) of anthropometric measurements on their own family, or on families with whom they are acquainted. The measurements are to be made use of for testing theories of heredity, no names, except that of the recorder, are required, but the Professor trusts to the *bona fides* of each recorder to send only correct results

Each family should consist of a father, mother, and at least one son or daughter, not necessarily the eldest. The sons or daughters are to be at least 18 years of age, and measurements are to be made on not more than two sons and two daughters of the same family. If more than two sons or daughters are easily accessible, then not the tallest but the eldest of those accessible should be selected.

To be of real service the whole series ought to contain 1000–2000 families, and therefore the Professor will be only too grateful if anyone will undertake several families for him.

The measurements required in the case of each individual are to be to the nearest quarter of an inch, and to consist of the following.

(I.) *Height* This measurement should be taken, if possible, with the person in stockings, if she or he is in boots it should be noted. The height is most easily measured by pressing a book with its pages in a *vertical plane* on the top of the head while the individual stands against a wall.

(II.) *Span of Arms* Greatest possible distance between the tip of one middle finger and the tip of the other middle finger, the individual standing upright against a wall with the feet well apart and the arms outstretched – if possible with one finger against a doorpost or corner of the room.

(III.) *The Length of LEFT Forearm* The arm being bent *as much as possible* is laid upon a table, with the hand flattened and pressed firmly against the table, a box, book, or other hard object is placed on its edge so as to touch the bony projection of the elbow, another so as to touch the tip of the middle finger. Care must be taken that the books are both perpendicular to the edge of the table. The distance between the books is measured with a tape.

Or,
The arm being bent *as much as possible* the elbow is pressed against the corner of a room or the doorpost, the hand being flattened and pressed against the wall. The greatest distance from the tip of the middle finger to the corner or doorpost is to be measured.

In this section, you will have the opportunity to explore the data that Pearson obtained on the heights of father–son pairs. One question we shall investigate is: 'Were the sons taller, on average, than the fathers?' That is, was the phenomenon of increasing stature, which has been observed more recently, evident in these families at the beginning of the 20th century? Another question is: 'Did tall fathers tend to have tall sons, and short fathers have short sons?' In this section, you will be able to investigate both these questions using the statistics software. We shall return to the second question in Chapter D4.

Activity 4.1 The data

Click on **Open...** in the **File** menu. The data on father son heights are contained in the file PEARSON.OUS. Open this data file now.

As you can see, the data are arranged in a frequency table, with the columns containing father's height in inches, son's height in inches, and frequency, respectively. You will see that, for example, there was one father–son pair with the father's height recorded as 59 inches and the son's height as 64 inches; and, if you scroll down to row 23, you will see that there were four father son pairs with the father's height 63 inches and the son's height 67 inches.

For paired data such as these, it is a good idea to begin by obtaining a scatterplot of the data. Click on **Plot** and choose **Scatterplot...** (by clicking on it). A 'Scatterplot' window appears. To obtain a scatterplot with father's height on the x-axis and son's height on the y-axis, select FatherHt | Frequency for the x variable and SonHt Frequency for the y variable. The scatterplot is then displayed lower in the window.

Is there any pattern discernible in the scatterplot? What does this tell you about the heights of the fathers and the sons?

Comment

The scatterplot is shown in Figure 4.1. (After choosing **Scatterplot...**, the size of the window was adjusted in order to obtain Figure 4.1.)

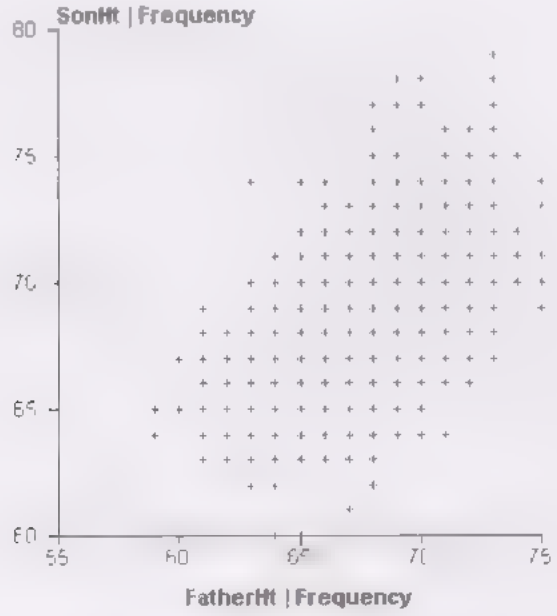


Figure 4.1 A scatterplot of son's height against father's height

Notice that some information is not shown in this scatterplot: it does not show how many father son pairs there were for each pair of heights, only whether or not there were any pairs. From the scatterplot, it appears that there is a tendency for the taller fathers to have sons taller than those of the shorter fathers. However, there is a lot of scatter, so the relationship between son's height and father's height is a weak one. It is not possible to tell from the scatterplot whether or not the average height of the sons is greater than the average height of the fathers.

Activity 4.2 Average heights

Use **Summary stats...** in the **Stats** menu to find the mean and standard deviation of the fathers' heights and of the sons' heights. (Select the variables FatherHt, Frequency and SonHt, Frequency to summarise the fathers' heights and sons' heights, respectively.)

Recall that you can select more than one variable at a time by holding down the `[Ctrl]` key while you click on variable names with the mouse.

Is the average height of the sons greater than the average height of the fathers? Are the sons' heights and the fathers' heights equally variable?

Comment

The sons were taller on average than the fathers – the mean height of the sons is 68.7 inches compared with 67.7 inches for the fathers. The heights of the fathers and the sons were equally variable – the standard deviations (2.75 inches for the sons and 2.72 inches for the fathers) are approximately equal.

Activity 4.3 Modelling the heights

(a) A frequency diagram of the sons' heights is shown in Figure 4.2(a). A normal curve with mean 68.7 and standard deviation 2.75 has been superimposed on the corresponding histogram in Figure 4.2(b).

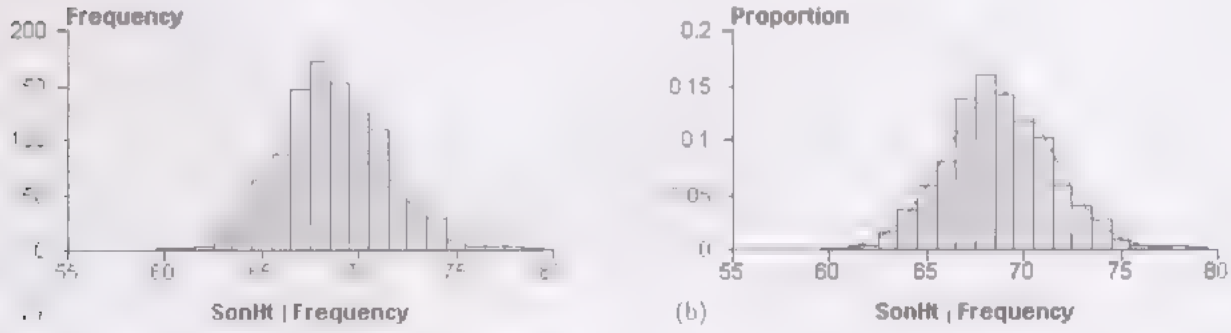


Figure 4.2 (a) A frequency diagram (b) The fitted normal curve

It looks as though a normal distribution models the variation in heights quite well.

Now follow the instructions below for using **Normal distribution...** to find the proportion of sons in this generation who were, according to this model, over six feet tall. This proportion is given by the area under the normal curve to the right of 72; this is shown in Figure 4.3.

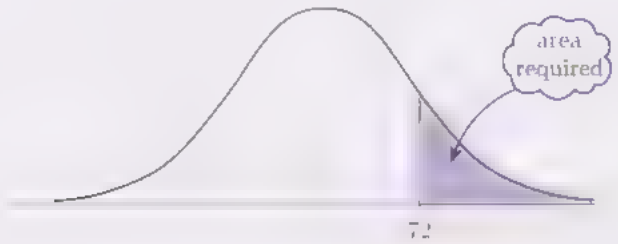


Figure 4.3 The area required

Select **Normal distribution...** from the **Stats** menu, and enter 68.7 and 2.75 for the mean and standard deviation, respectively. Click on **Update**, and the normal curve will be displayed.

First, find the area to the left of 72 as follows. Enter 72 in the **A** box and 72 (or any greater value) in the **B** box. Then click on the **Use A & B to calc areas** button (or [Tab] to this button and press the [Space] bar). The area is displayed in the **Area to left of A** box: it is 0.885 (to 3 significant figures). Since the total area under any normal curve is 1, the area required is equal to

$$1 - 0.885 = 0.115.$$

So, according to the model, approximately 11.5% of sons in this generation were over six feet tall.

(b) Now fit a normal distribution with mean 67.7 and standard deviation 2.72 to the data on the heights of the fathers. Use this model to estimate the proportion of fathers in this generation who were over six feet tall. Was there a greater proportion of sons over six feet tall than of fathers?

Comment

According to the model for fathers' heights, approximately 5.7% of fathers were over six feet tall. So a greater proportion of sons than fathers were over six feet tall.

In Activity 4.2, you found that the sons were taller, on average, than the fathers. And if anyone whose height is over six feet is regarded as tall, then it appears that there were more 'tall' sons than 'tall' fathers. However, to tackle the question of whether sons are taller than their fathers, we really need to look at the heights of sons whose fathers are of particular heights. For example, is the average height of the sons of fathers who were 64 inches tall greater than 64 inches? And is the average height of the sons of six-foot-tall fathers over six feet? You are asked to investigate questions such as these in the next two activities. The data are stored in a convenient form in the file SONS.OUS, so use this file for these activities. (The data in SONS.OUS and PEARSON.OUS are exactly the same; they are just arranged differently.)

Activity 4.4 Modelling sons' heights

In Activity 4.1, it was observed that sons of tall fathers tended to be taller than sons of short fathers. But if we know the height of a father, what precisely can we say about the height of his son? In this activity, you are invited to explore the heights of sons of fathers of various different heights.

(a) The data on the heights of sons of fathers who were 69 inches tall are contained in the columns named SonHt69 and Freq69 in the file SONS.OUS. Obtain a frequency diagram for these data.

You should find that it looks as though a normal distribution might provide a reasonable model for the variation in these heights. So fit a normal model to these data. (Remember to use one significant figure more for the parameters of the normal distribution than are given in the data: in this case, this means estimating the mean height to one decimal place and the standard deviation to two decimal places.)

According to the model, what proportion of the sons of 69-inch-tall fathers were more than 69 inches tall (and so taller than their fathers)?

(b) Investigate the heights of sons of fathers who were 71 inches tall. (The data are in the columns SonHt71 and Freq71.) Fit a normal distribution to the heights, and use it to estimate the proportion of sons of 71-inch-tall fathers who were taller than their fathers.

(c) Investigate the heights of sons of fathers who were 67 inches tall and of sons of fathers who were 64 inches tall. (These data are in the pairs of columns SonHt67, Freq67 and SonHt64, Freq64, respectively.) In each case, use a normal distribution to estimate the proportion of sons who were taller than their fathers.

Comment

(a) Using a normal distribution with mean 69.5 and standard deviation 2.30 in **Normal distribution...**, I found that the area to the left of 69 was equal to 0.414. So the proportion of sons over 69 inches tall was, according to the model, $1 - 0.414 = 0.586$, or approximately 59%.

(b) Using a normal distribution with mean 70.4 and standard deviation 2.48, I found that the area to the left of 71 was equal to 0.596. So the proportion of sons over 71 inches tall was, according to the model, $1 - 0.596 = 0.404$, or approximately 40%.

(c) Using a normal distribution with mean 68.0 and standard deviation 2.21, I found that the area to the left of 67 was equal to 0.325. So the proportion of sons over 67 inches tall was, according to the model, $1 - 0.325 = 0.675$, or approximately 68%.

Using a normal distribution with mean 66.6 and standard deviation 2.17, I found that the area to the left of 64 was equal to 0.115. So the proportion of sons over 64 inches tall was, according to the model, $1 - 0.115 = 0.885$, or approximately 89%.

It looks as though the sons of short men were more likely to be taller than their fathers than were the sons of tall men.

Activity 4.5 Looking for a relationship

In Activity 4.4, you found that, according to the models, although more than half the sons of fathers 64 inches, 67 inches or 69 inches tall were taller than their fathers, fewer than half the sons of fathers 71 inches tall were taller than their fathers. In this activity you are asked to investigate how the mean height of sons varies with father's height.

During Activity 4.4, you found some mean heights of sons for fathers of particular heights, as shown below.

Table 4.1

Father's height in inches	Mean height of sons in inches
64	66.6
67	68.0
69	69.5
71	70.4

The values in Table 4.1 were obtained from **Fit normal model**, but could also have been found from **Summary stats...**. The table can be extended by finding the mean height of sons in a similar way for every value of father's height from 59 to 75 inches. These data are contained in the file MEANS.OUS. Open this file now and check that the four data pairs in Table 4.1 agree with the corresponding rows of the data matrix.

You can see from the data in the columns FatherHt and SonMeanHt that the mean height of sons tends to increase with father's height. However, the sons of short fathers were not, on average, as short as their fathers; and the sons of tall fathers were not, on average, as tall as their fathers.

Obtain from the file MEANS.OUS a scatterplot, with father's height (FatherHt) on the x-axis and mean son's height (SonMeanHt) on the y-axis. What does the scatterplot tell you about the relationship between the heights of fathers and the mean height of their sons?

Comment

The scatterplot is shown in Figure 4.4.

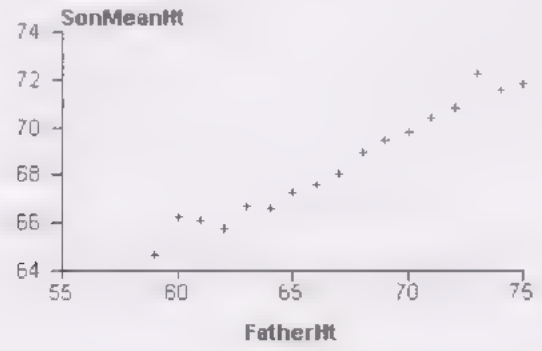


Figure 4.4 A scatterplot of mean son's height against father's height (heights in inches)

From the scatterplot you can see that the mean height of the sons tends to increase with father's height tall fathers tended to have tall sons, and short fathers tended to have short sons. However, an interesting point emerges from looking more closely at the data matrix and the scatterplot this is that, although tall fathers did tend to have tall sons, on average the height of the sons was less than the height of the fathers: for example, for fathers 73 inches tall, the average height of their sons was approximately 72.2 inches. Similarly, short fathers had sons who were not, on average, as short as themselves.

It was data similar to those collected by Pearson, but on a smaller scale (only about 200 father son pairs), that led Sir Francis Galton (1822 1911) in the 1880s to the idea of *regression*. Galton called the phenomenon just described 'regression back to the population mean' or, as he put it, 'toward the mediocre'. You can see from the scatterplot that the sample means lie approximately on a straight line, so it would seem reasonable to model the way the mean height of sons depends on father's height by a linear relationship. In Chapter D4, a method is described for choosing a line to model this relationship. The line obtained is called the *least squares fit line* or the *regression line*.

Obtaining a scatterplot: a summary

A scatterplot is obtained as follows.

- ◇ Choose **Scatterplot...** from the **Plot** menu.
- ◇ Select a variable name for the x variable and a variable name for the y variable. The scatterplot is then displayed.

Chapter D2, Section 5
Exploring normal distributions

If necessary, refer to page 17 for an introduction to use of the 'Normal distribution' window

In this section, you are invited to use *OUStats* to explore the properties of normal distributions. You will need to use **Normal distribution...** from the **Stats** menu for all the activities. In general, you will probably find it quicker and easier to edit values in the boxes than to use the mouse to mark areas under the normal curves.

Activity 5.1 Different means

For each pair of values of the parameters μ and σ in the table below, find the area under the normal curve between the values $a = \mu - \sigma$ and $b = \mu + \sigma$. What do you notice?

μ	σ	a $(= \mu - \sigma)$	b $(= \mu + \sigma)$	Area under curve between a and b
0	1	−1	1	

A solution is given on page 61.

Activity 5.2 Different standard deviations

(a) For each pair of values of the parameters μ and σ in the table below, find the area under the normal curve between the values $a = \mu - \sigma$ and $b = \mu + \sigma$. What do you notice?

μ	σ	a $(= \mu - \sigma)$	b $(= \mu + \sigma)$	Area under curve between a and b
0	1	−1	1	
0	5	−5	5	
0	140	−140	140	

(b) The results obtained in this activity and in Activity 5.1 illustrate a general result for normal distributions. Write down what you think this result might be. Test your conjectured 'result' for a pair of values of μ and σ of your own choice.

Solutions are given on page 61.

Activity 5.3 Two standard deviations from the mean

(a) For each pair of values of the parameters μ and σ in the table below, and for a pair of values of your own choice, find the area under the normal curve between the values $a = \mu - 2\sigma$ and $b = \mu + 2\sigma$. What do you notice?

μ	σ	a ($\mu - 2\sigma$)	b ($\mu + 2\sigma$)	Area under curve between a and b
0	1	2	2	
0	5	−10	10	
20	5	10	30	
20	50			

(b) These results illustrate a general result for normal distributions. Write down what you think this result might be. Test your conjectured 'result' for a pair of values of μ and σ of your own choice.

Solutions are given on page 61.

Activity 5.4 Three standard deviations from the mean

For each of the pairs of values of μ and σ in the table in Activity 5.3 (including those of your own choice), find the area under the normal curve between $\mu - 3\sigma$ and $\mu + 3\sigma$. Comment on your results.

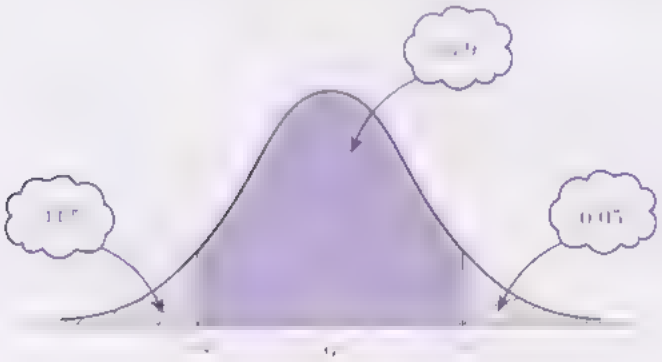
A solution is given on page 61.

Activity 5.5 90% of values

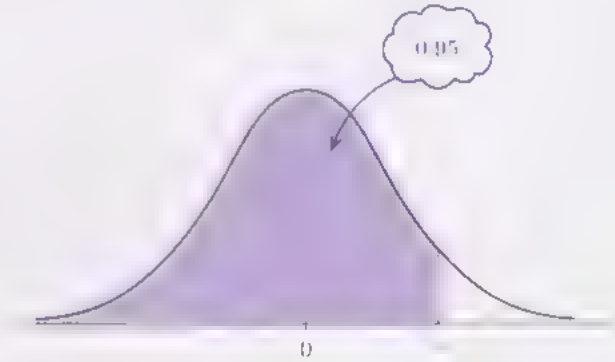
(a) Consider a normal distribution with parameters $\mu = 0$ and $\sigma = 1$. Follow the instructions below to find the value z such that the area under the normal curve between $-z$ and z is equal to 0.9.

To do this, you must first turn it into a question that you can answer using the **Normal distribution...** facility. The software allows you to find a value for **A** directly, given the value of **Area to left of A**, so the first thing to do is to calculate the area to the left of z.

The total area under the normal curve is 1, and the curve is symmetrical about the mean, $\mu = 0$; so, if the area between $-z$ and z is 0.9, then the area in each tail (that is, below $-z$ or above z) is equal to $\frac{1}{2} \times 0.1 = 0.05$. This is illustrated in Figure 5.1(a).



(a) The area between $-z$ and z



(b) The area to the left of z

Figure 5.1 Finding the area to the left of z

It follows that the total area to the left of z is 0.95; this is shown in Figure 5.1(b). So put 0.95 in the **Area to left of A** box, ensure that the value in the **Area between A and B** box is no greater than 0.05 and click on **Use areas to calc A & B**. The required value of z will appear in the **A** box.

If the area between A and B is ever as great as 0.5 then an error message will appear, since the sum of the two areas cannot exceed 1.

(b) Now consider a normal distribution with parameters $\mu = 20$ and $\sigma = 5$. Find the area under the normal curve between the values $\mu - z\sigma$ and $\mu + z\sigma$, where z is the value that you found in part (a).

(c) Select values for μ and σ different from those in parts (a) and (b). Find the area under the normal curve between the values $\mu - z\sigma$ and $\mu + z\sigma$, where z is the value that you found in part (a).

(d) Suggest a general result for normal distributions.

Solutions are given on page 61.

Activity 5.6 95% of values

(a) Consider a normal distribution with parameters $\mu = 0$ and $\sigma = 1$. Find the value z such that the area under the normal curve between $-z$ and z is equal to 0.95.

(b) Now consider a normal distribution with parameters $\mu = 20$ and $\sigma = 5$. Find the area under the normal curve between the values $\mu - z\sigma$ and $\mu + z\sigma$, where z is the value that you found in part (a).

(c) Select values for μ and σ different from those in parts (a) and (b). Find the area under the normal curve between the values $\mu - z\sigma$ and $\mu + z\sigma$, where z is the value that you found in part (a).

(d) Suggest a general result for normal distributions.

Solutions are given on page 61.

Activity 5.7 99% of values

(a) Consider a normal distribution with parameters $\mu = 0$ and $\sigma = 1$. Find the value z such that the area under the normal curve between $-z$ and z is equal to 0.99.

(b) Now consider a normal distribution with parameters $\mu = 20$ and $\sigma = 5$. Find the area under the normal curve between the values $\mu - z\sigma$ and $\mu + z\sigma$, where z is the value that you found in part (a).

(c) Select values for μ and σ different from those in parts (a) and (b). Find the area under the normal curve between the values $\mu - z\sigma$ and $\mu + z\sigma$, where z is the value that you found in part (a).

(d) Suggest a general result for normal distributions.

Solutions are given on page 62.

Chapter D3, Section 3
Confidence intervals on the computer

In Subsection 3.1, you will have the opportunity to use computer simulations to check the interpretation of a 95% confidence interval given in Section 2 of Chapter D3. And in Subsection 3.2, you will learn how to use *OUStats* to calculate a 95% confidence interval for a population mean from a large sample of data.

3.1 Interpreting a confidence interval

In Section 2, it was stated that if a large number of samples is drawn from a population, and a 95% confidence interval for the population mean is calculated from each sample, then approximately 95% of these confidence intervals will contain the population mean μ. In this subsection, you will have the opportunity to investigate the accuracy of this statement. You will be using the *Simulations* package to generate many samples and to calculate the corresponding confidence intervals.

For each different population distribution that you use, you can investigate what proportion of 95% confidence intervals contain the population mean, to ascertain whether it is indeed approximately 95%. The software can be used to simulate taking samples from either a normal distribution or a geometric distribution.

Activity 3.1 Confidence intervals for the mean of a normal distribution

(a) Start the *Simulations* package running as follows.

- Click on the **Start** menu, move the mouse pointer to **Programs**, then click on **MST121 Simulations**.

Alternatively, double-click on the **MST121 Simulations** [illegible]

Now click on **Confidence intervals** (either the tab or the panel) to open this simulation.

At the top of the window are two buttons, labelled **Normal** and **Geometric**. The default option is **Normal**; this is the option that you will be using in this activity.

The option of **Thick lines...** is available from the **Options** menu.

(b) The default values of the parameters μ and σ are 0 and 1, respectively; and the default values of the sample size and the number of samples are 25 and 100. Run the simulation with these values to see what happens.

Each confidence interval is represented by a horizontal line segment on the diagram. The population mean μ is marked by a vertical line down the centre of the diagram, and any confidence interval that does not contain μ is displayed in a different colour from those that do contain μ. Thus it is possible to identify easily those intervals that do not contain μ. When the simulation ends, you can scroll back to see how many of the intervals failed to include the population mean. Alternatively, notice that the number of confidence intervals which do contain the population mean μ is displayed in the box at the bottom left of the window.

There are also vertical lines at distances $\sigma/2$, σ and $3\sigma/2$ either side of the mean.

The first time that I ran the simulation, 94 out of the 100 confidence intervals contained the population mean μ; the other six did not. You may well have obtained a different number. Run the simulation several times, and note down the results. On average, approximately what proportion of your confidence intervals contained the population mean μ?

(c) Run the simulation several times for values of the parameters μ and σ of your own choice, and note down your results.

For each pair of values that you used, on average what proportion of the confidence intervals contained the population mean μ?

(d) Now change the sample size to 100 (say). Run the simulation several times for different values of μ and σ of your own choice, and note down your results in each case.

On average, what proportion of the confidence intervals contained the population mean μ?

(e) Investigate the proportion of confidence intervals which contain the population mean μ for further different sample sizes and parameter values. Note down your results.

(f) Comment briefly on your results and on any points you may have noticed about the confidence intervals that you obtained for different sample sizes.

Comment

I ran the simulation ten times for samples of size 25 from a normal distribution with mean 0 and standard deviation 1. I obtained the following results for the number of confidence intervals (out of 100) which contained the population mean μ.

94 92 93 97 94 97 95 97 91 94

Thus 94.4% of all the intervals contained the population mean μ. Another ten simulations produced 93.7% of intervals containing μ. In both cases, the proportion of confidence intervals which contained μ was only a little less than 95%.

For samples of size 100 and $\mu = 50$, $\sigma = 10$, I obtained the following results.

97 94 95 96 94 96 96 97 93 94

Overall, 95.2% of the confidence intervals contained the population mean μ. I used the simulation for various other sample sizes between 40 and 400, and for a number of different parameter values, and in each case obtained similar results: approximately 95% of the confidence intervals contained the population mean μ.

In general, I noticed that (as expected) the confidence intervals were narrower for larger sample sizes. It was observed earlier, in Chapter D2, that the sample standard deviation varies less from sample to sample when the sample size is large than when it is small. So we should expect the width of the confidence intervals to vary less from sample to sample for the larger sample sizes. This was so for my simulations. Did you notice that there was less variation in the width of the confidence intervals for the larger sample sizes that you tried than for the smaller sample sizes?

Activity 3.2 Confidence intervals for the mean of a geometric distribution

Now click on the **Geometric** button. The default value of the parameter p is 0.5.

Use this simulation to investigate the proportion of 95% confidence intervals for the mean of a geometric distribution that actually contain the population mean. Run the simulation for various values of the parameter p and for various sample sizes from 25 upwards.

Recall from Chapter D1 that a geometric distribution with parameter p has mean $\mu = 1/p$. It can also be shown that its standard deviation is $\sigma = \sqrt{1-p}/p$.

Comment briefly on your results. In particular, write down any points you may have noticed about the lengths of confidence intervals for different sample sizes, and about the proportion of confidence intervals which contain the population mean for different sample sizes.

Comment

I ran the simulation for a range of sample sizes similar to those which I used in Activity 3.1 when investigating confidence intervals for the mean of a normal distribution. I tried several values for the parameter p: $\frac{1}{6}, \frac{1}{2}, \frac{4}{5}, \ldots$. For each parameter value and for each sample size of 50 or larger that I tried, I found that the proportion of confidence intervals that contained the population mean was approximately 95%. However, when I took samples of size 25, the proportion of confidence intervals that contained the population mean was generally a little lower than 95%. For instance, for $n = 25$ and $p = \frac{1}{2}$, just over 91% of my intervals contained the population mean; and for $n = 25$ and $p = \frac{1}{6}$, only about 90% of my intervals contained the population mean. In all cases the proportion was less than 95%.

You may recall that, for large sample sizes, the sampling distribution of the mean may be approximated by a normal distribution, and the approximation improves as the size of the samples increases. A geometric distribution is right-skew (for any value of the parameter p), and the approximation is not nearly as good for samples of size 25 as it is for larger sample sizes. As a result, rather less than 95% of the confidence intervals actually contain the population mean for samples as small as 25.

This effect is seen at its most extreme for values of p close to 1. For example, with samples of size 25 and $p = 0.99$, less than 25% of confidence intervals contain the population mean. (Interpretation of the diagram in this case is an interesting exercise. The value of μ is $1/0.99 \simeq 1.01$, and many confidence intervals consist of the single value 1, corresponding to a sample of 25 values each of which is 1.)

The value of σ is $\sqrt{0.01}/0.99 \simeq 0.1$.

You may also have noticed that the width of the confidence intervals varied greatly for samples of size 25. This occurs because for samples of size 25 from a geometric distribution, the sample standard deviation varies greatly from sample to sample. As for the normal distribution, the sample standard deviation, and hence the width of the confidence intervals, varies much less from sample to sample for larger sample sizes.

3.2 Calculating a confidence interval

In order to calculate a 95% confidence interval for a population mean, the values of the sample mean $\bar{x}$ and the sample standard deviation s need to be calculated. For large samples, this can be a tedious exercise using a calculator, so you were spared carrying out these calculations in Section 2; in each example and activity, you were given the values of $\bar{x}$ and s.

In this subsection, you will not be given these summary statistics. Instead, you will be given the data in a file, and invited to use *OUStats* to calculate confidence intervals. The sample mean and sample standard deviation are calculated automatically when using *OUStats* to find a confidence interval.

In the first activity, you will be shown how to find a confidence interval for the mean height of Cambridge men in 1902. You will be able to check that the calculations agree with those carried out using a calculator in Section 2. You will need to use *OUStats* for all the activities in this subsection.

To start up *OUStats*, click in turn on **Start**, **Programs** and **MST121 OUStats**, or just double-click on the **MST121 OUStats** desktop icon.

Activity 3.3 Finding a confidence interval

The data on the heights of 1000 Cambridge men are contained in the file HEIGHTS.OUS. Open this file now.

A 95% confidence interval for a population mean is obtained using **Confidence interval...** in the **Stats** menu. Click on **Stats**, and choose **Confidence interval...** (by clicking on it). From the list of variables that appears, select 'Height | Frequency', and click on **Select**. The confidence interval is then calculated.

The output includes the sample mean, sample standard deviation and sample size (for information), and a statement of the 95% confidence interval for the population mean. In this case, the confidence interval given is (68.7128, 69.0312). So we can be fairly sure that the mean height of all Cambridge men in 1902 was between 68.71 inches and 69.03 inches.

In Section 2, we obtained (68.7, 69.1) for the 95% confidence interval. The slight discrepancy between these two results is due to rounding error: in Section 2, we used values for the mean and standard deviation which had been rounded to 3 significant figures, whereas *OUStats* calculates the values of the mean and standard deviation to much greater accuracy and uses these values to calculate a confidence interval.

Activity 3.4 Sample size

The procedure described in Section 2 for calculating a 95% confidence interval for a population mean should be used only when the sample size is at least 25. Any results obtained using the formula given there for a smaller sample would be inaccurate and unreliable. In this activity, you are asked to explore what happens if you try to use *OUStats* to calculate a confidence interval for a sample of fewer than 25 items of data.

The data file BAR.OUS contains data on the gross hourly earnings (in pence) in 1995 of a sample of 14 female bar staff. Open the file now, and instruct the computer to calculate a 95% confidence interval for the mean gross hourly earnings in 1995 for female bar staff. What output do you obtain?

Comment

You will have found that, because the sample size is less than 25, the software produces a message telling you that the sample size must be at least 25. However, most statistics packages are not so friendly: if you ask for an inappropriate procedure to be carried out, it will be done. So it is important for you to know when a procedure should or should not be used.

Activity 3.5 Cuckoo eggs

The lengths in millimetres of the 243 cuckoo eggs which were represented in Figure 1.5(d) of Chapter D2 are contained in the file CUCKOOS.OUS. Use these data to find a 95% confidence interval for the mean length of all cuckoo eggs.

Source: O. H. Latter, 'The egg of *Cuculus canorus*', *Biometrika* 1 (1902) pages 164–176.

A solution is given on page 62.

Activity 3.6 Authorship and sentence length

In Activity 2.4, you calculated a 95% confidence interval for the mean sentence length in a book by G. K. Chesterton. The data on sentence lengths are contained in the data file AUTHORS.OUS, together with data on sentence lengths in two other books: *The Work, Wealth and Happiness of Mankind* by H. G. Wells and *An Intelligent Woman's Guide to Socialism* by G. B. Shaw. These data were collected by C. B. Williams in an investigation into sentence length as a criterion of literary style. (Williams chose these particular books for his investigation because, in his words, 'all three deal with sociological subjects and none of them are in the "conversational style"'.)

Source: C. B. Williams, 'A note on the statistical analysis of sentence-length, as a criterion of literary style', *Biometrika* 31 (1940) pages 356–361.

Open the file AUTHORS.OUS now, and explore its contents: remember that you can obtain information about the data in the file using **Notes...** from the **File** menu.

(a) How does the distribution of sentence lengths vary between authors? In order to help you answer this question, obtain frequency diagrams for each of the authors, and compare them.

Hint: Choose **Frequency diagram...** from the **Plot** menu to obtain each of the diagrams in turn. Then choose **Tile** from the **Window** menu, so that you can view all three diagrams together. The scales on the three diagrams will be different.

If you wish, you can adjust the size of a window, and hence the appearance of any scale on a diagram within, by dragging the mouse. Instructions for doing so were given just before Activity 3.3 of Chapter D2 (page 27 in this book).

(b) Compare the mean sentence lengths for the three authors. Does there appear to be a difference? Which author seems to write the longest sentences? Which author seems to write the shortest sentences?

(c) Find a 95% confidence interval for the mean sentence length in each book. What do you conclude from your results?

Solutions are given on page 62.

Activity 3.7 Birthweights

The file BIRTHWT.OUS contains the birthweights of 37 male and 34 female babies, all of whom were born two weeks 'early', that is, at the end of a 38-week gestation period. Find 95% confidence intervals for the mean birthweight of baby boys born two weeks early and for the mean birthweight of baby girls born two weeks early. Comment on your results.

A solution is given on page 63.

Calculating confidence intervals: a summary

A 95% confidence interval for a population mean based on a large (≥ 25) sample from the population is obtained as follows.

- ◇ Choose **Confidence interval...** from the **Stats** menu.
- ◇ Select the appropriate variable name(s) from the variable list that appears (to indicate the data to be used), and click on **Select**.

A 95% confidence interval is calculated using each set of data selected.

Chapter D4, Section 2
Exploring the data

In this section, the use of *OUStats* to produce boxplots is illustrated for the data on city block scores which are given in Table 1.1 of Chapter D4. You will be invited to explore the data further to see whether there is a relationship between the time spent memorising the positions of the objects and the score obtained on the test

Activity 2.1 Obtaining boxplots

The data on city block scores and on memorisation times are contained in the file MEMORY.OUS. Open the file now, and read the information about the data given in **Notes...**.

Boxplots are obtained using the **Plot** menu. Click on **Plot**, and choose **Boxplot...** (by clicking on it). To obtain boxplots for the city block scores of the two groups on the same diagram, select YScore and EScore as follows: click on YScore, then, while holding down the `[Ctrl]` key, click on EScore; both YScore and EScore should now be highlighted. Click on **Select**, and the boxplots will be produced.

You should find that the boxplots look similar to those in Figure 1.4 of Chapter D4, although the five key values are not displayed. When a 'Boxplot' window is open, you can display a list of the five key values by clicking the mouse button while the pointer is within the box or close to either whisker. Try this now.

Next obtain boxplots for the memorisation times of the two groups, YTime and ETime (on a single diagram). Check that they look similar to those in the solution to Activity 1.5(b) of Chapter D4.

In Section 1, we observed that the boxplots for the city block scores suggest that, generally, the young people performed better on the test than the elderly people. However, the boxplots for the memorisation times indicate that the young people spent longer studying the positions of the objects. Does spending longer studying the positions improve performance on the test? If so, then this could explain why the young people performed better on the test.

To investigate whether memorisation time and performance on the test are related, we must look at the city block scores and memorisation times of the individuals who took the test. This information is available in the file MEMORY.OUS. The data in the file are paired. For instance, the first entry in the column headed EScore and the first entry in the column headed ETime relate to one person from the elderly group, and so on.

Activity 2.2 Is performance related to memorisation time?

Instructions for obtaining a scatterplot were given in Chapter D2, Activity 4.1 (page 32 in this book).

Obtain two scatterplots, one for the young people and one for the elderly people. Plot memorisation time on the horizontal axis, and city block score on the vertical axis. (Use **Scatterplot...** from the **Plot** menu.) Is there any evidence of a relationship between memorisation time and city block score for either group? Describe any patterns in the scatterplots.

Comment

You should find that, in both scatterplots, there is a tendency for the city block score to decrease as the memorisation time increases. However, there is a lot of scatter in the plots, so the relationships are weak.

Activity 2.3 Combining the data

It is difficult to tell from the two separate scatterplots whether a young person and an elderly person who spend similar times studying the positions of the objects obtain similar city block scores. We can investigate this by plotting all the data on the same diagram. Obtain a scatterplot for all 27 people who took the test, with memorisation time on the horizontal axis and city block score on the vertical axis. The data for all 27 people are in the columns headed Score and Time.

What can you deduce from the scatterplot? Is there any evidence that young people do better on the test — that is, have lower city block scores than elderly people who spend a similar length of time memorising the positions of the objects?

Comment

Figure 2.1 contains a scatterplot showing the city block scores and memorisation times for all 27 people who took the test. Different plotting symbols have been used in this figure for the young and the elderly. (It is not possible to use different symbols using *OUStats*.)

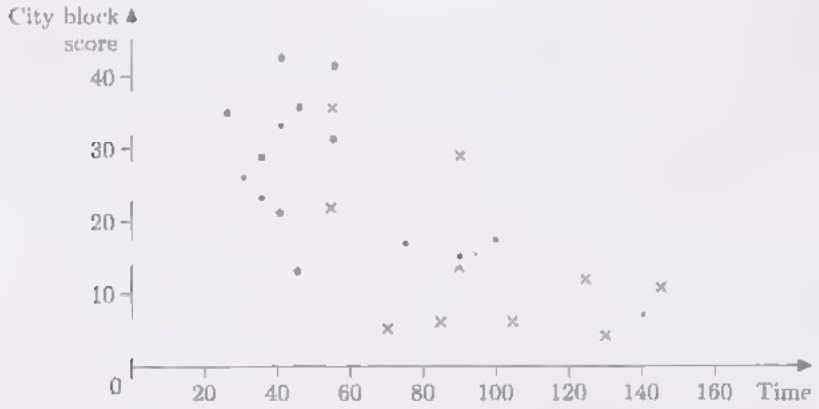


Figure 2.1 A scatterplot of city block scores and memorisation times (× for Young, • for Elderly)

Looking at the scatterplot as a whole, there appears to be a relationship between the time spent studying the positions of the objects and the city block score obtained: in general, the city block score decreases as the memorisation time increases. There is some overlap in the memorisation times of the people in the young and the elderly groups: in each group, there were individuals who spent between 55 and 100 seconds studying the positions of the objects. This part of the scatterplot provides very little evidence that city block scores are lower for young people than for elderly people who spent similar times memorising the positions of the objects.

It seems possible that the better performance of the young group on the test is due to the fact that, in general, they spent longer than the elderly group studying the positions of the objects. Of course, we do not know whether the elderly people would have done as well as the young people if they had all spent the same time studying the positions of the objects.

This will be investigated further in Section 6.

The final activity in this section will give you some practice at obtaining boxplots on the computer and interpreting them. You will obtain further practice in Section 4.

Activity 2.4 Earnings of primary school teachers

The file PRIMARY.OUS contains data on the gross weekly earnings in 1995 of 91 primary school teachers, of whom 54 are women and 37 are men. Obtain boxplots for the earnings of the women and the men. What do the boxplots tell you about the relative earnings in 1995 of male and female primary school teachers?

A solution is given on page 63.

Obtaining boxplots: a summary

Boxplots are obtained as follows.

- Choose **Boxplot...** from the **Plot** menu.
- Select one or more variable names, and click on **Select**.
- To display a list of the five key values on a boxplot, click the mouse button while the pointer is within the box or close to one of the whiskers.

Chapter D4, Section 4
Testing for a difference

In this section, the use of *OUStats* to carry out a two-sample z-test is explained. An essential first step in any investigation is to look at the data. So, in each case, you will be asked to compare the data visually using boxplots before performing a two-sample z-test.

Activity 4.1 Wing lengths of meadow pipits

In this activity, the data from Table 3.1 of Chapter D4 on the wing lengths of male and female meadow pipits will be used to demonstrate the use of *OUStats* to perform a two-sample z-test.

(a) The data are in the file PIPITS.OUS. Open this file now. Compare the wing lengths of the male and female meadow pipits using boxplots. Check that these boxplots agree with those given in Figure 3.1 of the main text in Chapter D4.

The boxplots suggest that there is a difference between the wing lengths of male and female meadow pipits. So now we shall carry out a two-sample z-test to investigate this apparent difference. The first stage is to write down the null and alternative hypotheses: these are

$$H_0 : \mu_M = \mu_F,$$
$$H_1 : \mu_M \neq \mu_F,$$

where μ_M, μ_F are the mean wing lengths of the populations of male and female meadow pipits, respectively. (These hypotheses were stated in Section 3.)

The second stage is to calculate the test statistic. This is the part of the hypothesis test that the computer can do for you. Click on **Stats**, and choose **Two sample z-test...** (by clicking on it). You need to specify the data to be used. Select MLength (which contains the wing lengths of the males) as the first variable, and FLength (which contains the wing lengths of the females) as the second variable. Click on **Calc**, and the calculations will be performed.

The output includes the mean, the standard deviation and the sample size of each of the two samples, and the test statistic. According to *OUStats*, the numerical value of z, the test statistic, is 7.5624. Notice that this differs slightly from the value we obtained in Section 3: there we obtained $z = 7.63$. This discrepancy is due to rounding error: in Section 3, to calculate the test statistic, we used values of the means and standard deviations which had been rounded to three significant figures, whereas *OUStats* calculates the means and standard deviations to many more significant figures than this, and then uses these values to calculate the test statistic.

The third and final stage in a hypothesis test is to draw a conclusion (as in Section 3). Since the test statistic z equals 7.5624, which is greater than 1.96, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that the

mean wing length of male meadow pipits is not equal to the mean wing length of female meadow pipits. And since the sample mean is greater for the males than for the females, this suggests that the mean wing length of males is greater than the mean wing length of females.

Activity 4.2 Sample sizes

In Section 1, the city block scores on a memory test of a group of 13 young people and 14 elderly people were compared using boxplots. And in Section 2, you reproduced these boxplots using *OUStats*. Since the two-sample z-test depends on the Central Limit Theorem, both sample sizes must be at least 25 for the test to be used. Try using the software to perform the test for the data on city block scores. (The data are in the file MEMORY.OUS.) What happens?

Comment

You will have found that, for these data, the software produces a message telling you the sample sizes and reminding you that both sample sizes must be at least 25. The test is not carried out. This happens whenever either of the sample sizes is less than 25. This is another friendly feature of the software: most statistics packages will carry out your instructions to perform a two-sample z-test whether or not it is an appropriate procedure to use. If it is not appropriate, because the sample sizes are too small, then the results produced using the test would be unreliable and possibly misleading.

Activity 4.3 Authorship and sentence length

In Activity 3.6 of Chapter D3 (page 45 in this book), you explored the distribution of sentence lengths for three books, one by each of three authors – G. K. Chesterton, H. G. Wells and G. B. Shaw. You also obtained confidence intervals for the mean sentence lengths of the three books, and compared them. In this chapter, two methods for comparing samples of data have been described: first boxplots for a visual comparison, and then the two-sample z-test to test for a difference between two population means.

(a) Compare the sentence lengths of the three authors using boxplots (the data are in the file AUTHORS.OUS).

(b) Use the two-sample z-test to investigate whether there is a difference between the mean sentence lengths in the book by G. K. Chesterton and the book by H. G. Wells. State your hypotheses, the test statistic and your conclusion clearly. Note that you should include a statement of your hypotheses, the test statistic and your conclusions in your record of every hypothesis test that you carry out. This applies whether you use a calculator or a computer to carry out the calculations.

(c) Test for a difference between the mean sentence lengths in the book by G. K. Chesterton and the book by G. B. Shaw. Again, state clearly your hypotheses, the test statistic and your conclusion.

Solutions are given on page 64.

Activity 4 4 Birthweights of babies

The file BIRTHWT.OUS contains the birthweights of 37 male and 34 female babies, all of whom were born two weeks 'early', that is, at the end of a 38-week gestation period. In Activity 3.7 of Chapter D3 (page 46 in this book), you found 95% confidence intervals for the mean birthweight of baby boys born two weeks early and for the mean birthweight of baby girls born two weeks early.

(a) Obtain boxplots for the birthweights of the boys and girls. Comment on what they tell you about the birthweights of boys and girls born two weeks early.

(b) Use the two-sample z-test to investigate whether there is a difference between the mean birthweight of boys born two weeks early and the mean birthweight of girls born two weeks early. State clearly your hypotheses, the test statistic and your conclusion.

Solutions are given on page 64.

Activity 4 5 Earnings of primary school teachers

The file PRIMARY.OUS contains data on the gross weekly earnings in 1995 of 91 primary school teachers, of whom 54 are women and 37 are men. In Activity 2.4 (page 49 in this book), you compared the earnings of the men and women using boxplots. Use the two-sample z-test to investigate whether there was a difference between the mean gross weekly earnings (in 1995) of male primary school teachers and female primary school teachers. State clearly your hypotheses, the test statistic and your conclusion.

A solution is given on page 65.

Two-sample z-test: a summary

The two-sample z-test is carried out as follows.

- Choose **Two sample z-test...** from the **Stats** menu.
- Select two variables, one from each of the two drop-down menus, and click on **Calc**.

The test is carried out only if both sample sizes are at least 25. If either sample size is less than 25, then an error message is produced.

Chapter D4, Section 6
Fitting a line to data

In this section, *OUStats* will be used to calculate the least squares fit line for the concrete data given in Section 5. You will then have the opportunity to investigate the relationships between several other pairs of variables. In each case, you will be asked to obtain a scatterplot and, if it seems appropriate, to fit a straight line to the data and use the equation of this line to make predictions.

Activity 6.1 Finding the least squares fit line

The data on the pulse velocity and crushing strength of concrete, given in Table 5.1 of Chapter D4, are contained in the file CONCRETE.OUS.

(a) Open the file now, and obtain a scatterplot of the data with pulse velocity along the x-axis and crushing strength along the y-axis.

You can now display the regression line on the scatterplot by opening the drop-down **Options** menu within the 'Scatterplot' window and clicking on 'Regression line on/off'. The resulting display also shows the parameters of the regression line: a slope of 25.887 42 and a y-intercept of $-87.828\,33$. (The regression line may be removed from the scatterplot by repeating these steps.)

This drop-down menu also provides the opportunity to alter the symbol used for plotting data points

(b) The equation of the least squares fit line can also be obtained using the **Stats** menu. Choose **Regression...** (by clicking on it). Now select 'Velocity' as the first (x) variable and 'Strength' as the second (y) variable, and click on **Calc**. The equation of the least squares fit line is displayed lower in the window, in the form

$$y = -87.8283 + 25.8874x.$$

So the equation of the least squares fit line is $y = -87.83 + 25.89x$, where y is the crushing strength of concrete and x is the pulse velocity for the concrete. This is the equation that was quoted in Section 5.

Activity 6.2 How tall will my son be?

Pearson's data on the heights of 1078 father–son pairs are contained in the file PEARSON.OUS.

(a) Obtain a scatterplot of son's height against father's height (with father's height along the x-axis), and then add the least squares fit line to the plot.

(b) Obtain the equation of the least squares fit line, and use it to predict the height of the son of a 70-inch-tall man.

(c) By referring to the scatterplot you obtained in part (a), comment on how precise you think this estimate might be.

Solutions are given on page 65.

When will Old Faithful erupt?

See Chapter D2, Activity 1.2.

Every year, tourists flock to the Yellowstone National Park in Wyoming in the United States. One of the attractions is the Old Faithful geyser, which erupts about 20 times a day, on average. As you saw in Chapter D2, the eruptions vary in length, the shortest lasting just over a minute and the longest about 5 minutes. The intervals between eruptions also vary a lot. Sometimes the waiting time from the end of one eruption to the beginning of the next is as short as 40 minutes, but it can be as long as an hour and a half. Unlucky visitors can have a long wait! So is there a way of predicting when the next eruption will occur, so that visitors can be informed?

In August 1978, the geyser was observed between 6 am and midnight on eight consecutive days; the duration of each eruption and the waiting time until the next eruption were both recorded. The purpose of collecting the data was to investigate whether the duration of one eruption could be used to predict when the next is likely to occur. The question was: 'Is there a relationship between the duration of an eruption and the waiting time until the next eruption?' And if there is, can we use the data to formulate a rule for predicting when the next eruption is likely to occur?

Activity 6.3 Exploring the relationship

The data on the eruptions of the Old Faithful geyser in August 1978 are contained in the file FAITHFUL.OUS.

(a) Obtain a scatterplot with the duration of an eruption along the x-axis and the waiting time until the next eruption along the y-axis.

(b) What does the scatterplot tell you about the relationship between the duration of an eruption and the waiting time until the next eruption? Do you think a straight line would be a suitable model for the relationship?

(c) Obtain the equation of the least squares fit line, and use it to predict the waiting time until the next eruption following eruptions which last for the following times.

(i) 1.5 minutes (ii) 3 minutes (iii) 4.5 minutes

(d) Add the least squares fit line to your scatterplot (if not already done). Comment on how accurate you think your predictions are.

Solutions are given on page 66.

In fact, when the data were collected, an error was made in recording one day's results: the eruption times and intervals between eruptions were paired incorrectly. This meant that there were quite a number of anomalous points which did not fit the general pattern that you observed. So it was thought that a useful prediction rule could not be formulated. The error was discovered only several years later.

Memory and age

In Section 1 of Chapter D4, an investigation into spatial memory in the young and elderly was discussed. Two groups of people, one young and one elderly, tackled a memory test in which eighteen everyday objects were placed on a 10 by 10 square grid. After a person had studied the positions of the objects for as long as they wished, the objects were removed. Then they were asked to replace the objects in the same positions. Two pieces of data were noted for each person: the time spent studying the positions of the objects, and a measure of accuracy of recall — the city block score.

The city block score is described in Chapter D4, Activity 1.

Activity 6.4 Does performance improve with time?

In Section 2 of Chapter D4, you used *OUStats* to investigate whether performance on the memory test was related to the time spent memorising the positions of the objects. The data are in the file MEMORY.OUS.

(a) For the elderly group, obtain a scatterplot of city block score against memorisation time. Is there any evidence of a relationship between the time spent memorising the positions of the objects and performance on the test? If you think it appropriate, fit a line to the data (with [illegible]).

You obtained the three scatterplots required in Activity 6.4 previously, in Activities 2.2 and 2.3. However, the [illegible] fitting a line to the data do not form part of those activities.

(b) Now obtain a similar scatterplot for the young group. Comment on the relationship between the time spent memorising the positions of the objects and performance on the test. If you think it appropriate, fit a line to the data.

(c) Now obtain a scatterplot for the data for the two groups combined, and fit a line to the data. According to this model, what is the predicted score of a person whose time spent memorising the positions of the objects is as follows?

Note that the memorisation times are recorded in seconds.

(i) 1 minute (ii) 2 minutes (iii) 3 minutes

Comment briefly on your results.

Solutions are given on page 66.

Activity 6.5 Comparing the fit lines

Obtain a printout of the scatterplot for the two groups combined (without the least squares fit line on it). On this scatterplot, draw the two fit lines whose equations you found in parts (a) and (b) of Activity 6.4 — one for the elderly group and one for the young group. Comment briefly on what you deduce from this diagram.

A solution is given on page 67.

Finding the equation of the regression line: a summary

The equation of the least squares fit line, or the regression line of y on x, can be obtained as follows.

- ◇ Choose **Regression...** from the **Stats** menu.
- ◇ Select a variable name for the first (x) variable and a variable name for the second (y) variable, and click on **Calc**.

Alternatively, the parameters of the equation are given along with the least squares fit line that can be added to a scatterplot, as below.

Scatterplots: a summary

A scatterplot is obtained as follows.

- ◇ Choose **Scatterplot...** from the **Plot** menu.
- ◇ Select a variable name for the x variable and a variable name for the y variable. The scatterplot is then displayed.

The least squares fit line may be included on a scatterplot as follows.

- ◇ Open the **Options** drop-down menu within the 'Scatterplot' window.
- ◇ Choose 'Regression line on/off' from the list of options (by clicking on it). The regression line appears, along with the values of the parameters for its equation.

The regression line may be removed by repeating these steps.

The plotting symbol on a scatterplot may be changed as follows.

- ◇ Open the **Options** drop-down menu within the 'Scatterplot' window.
- ◇ Choose the symbol that you require from the list that appears (by clicking on the corresponding 'Plot points as' row).

Appendix: Entering and editing data

Creating a new data file

- Open *OUStats*. Choose **New...** from the **File** menu. You will then see a dialogue box for entering the numbers of rows and columns that you want to have available. The minimum size of grid is 200 rows by 20 columns, and this may suffice for most purposes. If you need more rows or columns than this, edit the appropriate box. Then click on **OK** or press [Enter].
- You will then see a data window showing a data matrix of the size that you specified in the dialogue box. To enter a column of data in, say, column V1, click on the first row of V1 and type in the first value. Press [Enter] to move the cursor to the next row, and type in the next value. Press [Enter] again, and continue entering values in this way. To type in a second column of data, click on the first row of that column and repeat the entering of values as for the first column. Note that if you prefer to enter data across the rows rather than down the columns, you can simply press [Tab] instead of [Enter] after typing each value. This will move the cursor to the right.
- You can go back and edit any value by clicking on it, using [Backspace] to delete, typing in the new value, and pressing [Enter]. You can also move around the data matrix using the arrow keys.

Naming and renaming columns

When a new data window is selected, the default names for the columns (that is, variables) are V1, V2, V3, etc. To rename a column (once some data have been entered), choose **Rename column...** from the **Edit** menu, select the original name in the dialogue box which appears (by clicking on the corresponding editing box), type in the new name, and click on **OK**.

Saving a new data file

Choose **Save As...** from the **File** menu. You should then see a 'Save' dialogue box which shows a scrollable list of data (.OUS) file names.

Enter a name for your data file. Then click on **Save**. Your file is now saved in the same folder as all the other data files, and should be listed with them when you choose **Open...** from the **File** menu.

Entering frequency data

Frequency data must be entered with the values in one column (in V1, say) and the corresponding frequencies in another column (in V2, say). To designate the values in a column as values for frequency data, place the mouse pointer in the heading for that column (V1 in this case) and click with the *right* mouse button. Then choose (for this example) 'Set frequency: V2' from the menu that appears. The values in column V2 are then designated as frequencies for the values in column V1. (You can check this by clicking again with the right mouse button in the heading for column V1. You will notice that the top (highlighted) line of the resulting menu now reads 'Frequency: V2' in place of 'No frequencies'.)

You must enter all values in the columns and, if needed, rename the columns, before designating values as frequency data.

When the values in one column have been designated in this way as frequencies for another column, the names of the two columns are linked together in dialogue boxes which display variable names. For instance, 'Height | Frequency' refers to the frequency data which have been entered into the two columns labelled 'Height' and 'Frequency', then linked together as described above.

Making your own Notes file

When you create a new data file using *OUStats*, it is a good idea to record information about the file (the source of the data, an explanation of the variable names, etc.) in an associated Notes file. You can do this as follows.

- Open the data file (if it is not already open).
- Click on **Notes...** in the **File** menu.
- Type your notes about the data file into the 'Notes' window.
- To save the notes, click on **Save** in the **File** menu (whether or not the 'Notes' window itself is open).

The commands **Cut**, **Copy** and **Paste**, in the **Edit** menu, are available whenever a 'Notes' window is open.

Transferring data into and out of OUStats

Instructions for transferring data from another application into *OUStats*, or vice versa, are given in the *OUStats* Help file, which may be accessed via the *OUStats* **Help** menu.

Chapter D2

Solution 3.1

Frequency diagrams for the four data sets are shown in Figure S2.1.

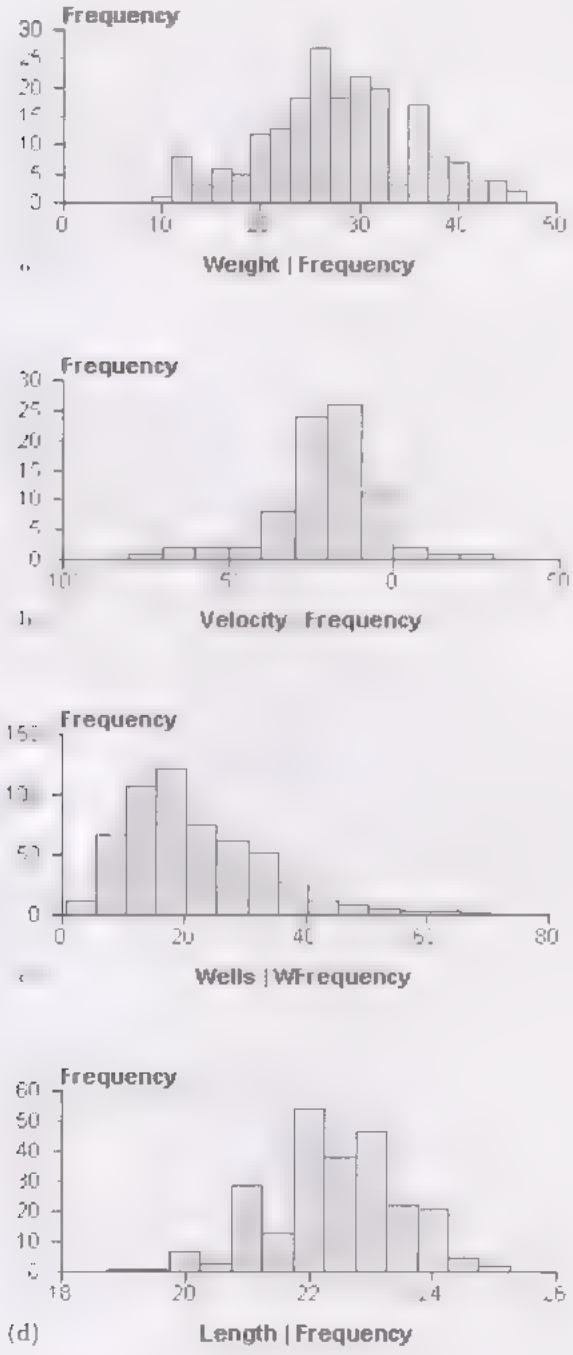


Figure S2.1 Four frequency diagrams

(a) The first frequency diagram (for the weights of Irish dipper nestlings) was obtained using the first interval starting value and interval width suggested in the activity (9 and 2, respectively).

(b) For the frequency diagram for the radial velocities, a first interval starting value of −80 and an interval width of 10 were used, corresponding to the grouping of the data.

(c) To obtain the frequency diagram for the lengths of sentences written by H. G. Wells, a first interval starting value of 0.5 and a width of 5 were used. This means that the first bar represents sentences of lengths 1 to 5 inclusive, the second bar represents sentences of lengths 6 to 10 inclusive, and so on. You may well have chosen different values for the start of the first interval and the interval width.

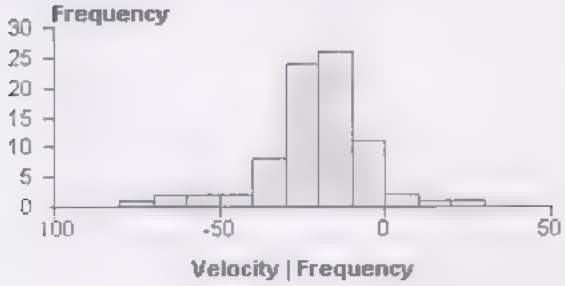
(d) The lengths of cuckoo eggs are given to the nearest half millimetre, so a length recorded as 19 mm could be anywhere between 18.75 mm and 19.25 mm. To obtain the frequency diagram shown, a first interval starting value of 18.75 and an interval width of 0.5 were used.

The frequency diagram for the lengths of sentences written by H. G. Wells is right-skew (the right tail is longer than the left tail), so a normal distribution is not an appropriate model in this case. The other three frequency diagrams are all roughly symmetrical with a single clear peak, so a normal model is worth considering for these data. You are asked to investigate these three data sets further in the next three activities.

Solution 3.3

I fitted a normal model with mean −21 and standard deviation 16. (As in Activity 3.2, I used one significant figure more for the parameters of the normal distribution than are given in the data.)

The four frequency diagrams obtained are shown in Figure S2.2. (The size of each random sample was 80, the same as the size of the sample of data.)



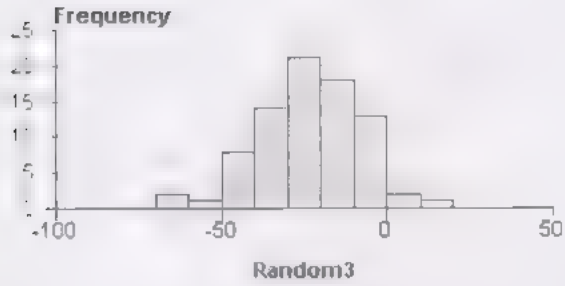


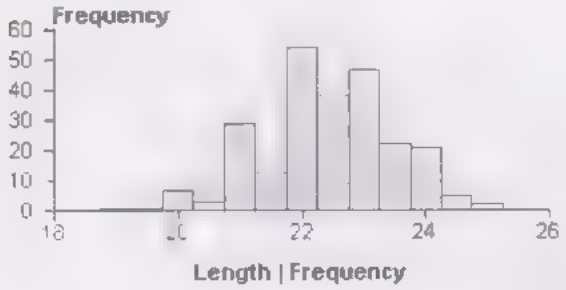
Figure S2.2 Four frequency diagrams

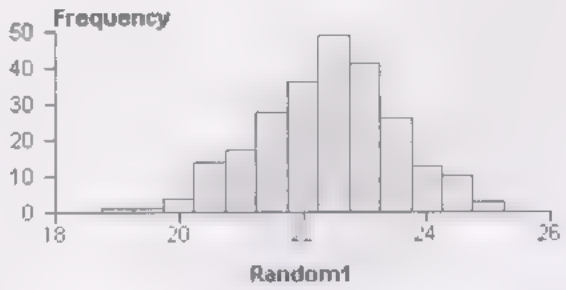
There is considerable variation between the shapes of the frequency diagrams for the random samples. However, they all appear to be less sharply peaked than the frequency diagram for the data, which has longer shallower tails. It would appear that a normal model may not be a very good fit for the data. Certainly, the fit does not seem to be as good in this case as the fit of the normal curve to the weights of Irish dipper nestlings in Activity 3.2.

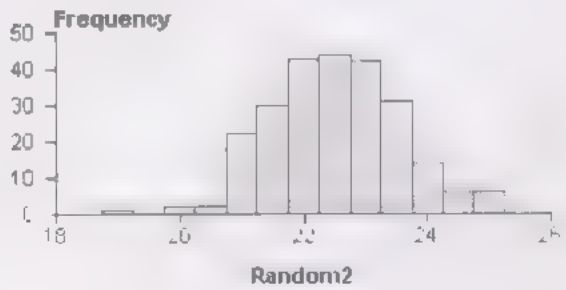
Solution 3.4

I fitted a normal distribution with mean 22.4 and standard deviation 1.08. (Again, I used one significant figure more for the parameters of the normal distribution than are given in the data.)

The four frequency diagrams obtained are shown in Figure S2.3. (The size of each random sample was 243, the same as the size of the sample of data.)







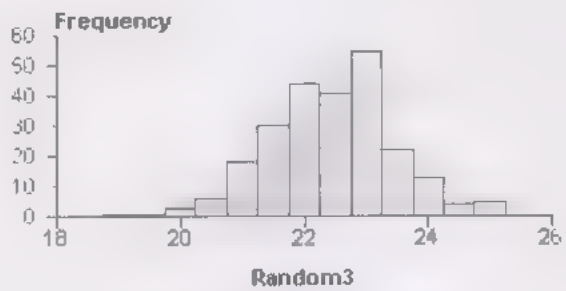


Figure S2.3 Four frequency diagrams

Again, there is considerable variation between the shapes of the frequency diagrams for the random samples. The main difference between the frequency diagram for the data and the frequency diagrams for the random samples is that the frequency diagram for the data is more jagged. Apart from this, the fit seems to be quite good.

Solution 5.1

All the areas between a and b are equal to 0.683 (to 3 s.f.).

Solution 5.2

(a) All the areas between a and b are equal to 0.683 (to 3 s.f.).

(b) The area under a normal curve from one standard deviation below the mean to one standard deviation above the mean is the same whatever the mean and standard deviation of the distribution. In fact, the proportion of values within one standard deviation of the mean is about 68.3% for any normal distribution.

Solution 5.3

(a) All the areas between a and b are equal to 0.954 (to 3 s.f.).

If you have *OUStats* set to display 6 significant figures (the default), then it will show these areas as 0.9545 (suppressing two trailing zeros). However, to 8 s.f. the areas are displayed as 0.954 499 74, giving 0.954 to 3 s.f.

(b) The area under a normal curve from two standard deviations below the mean to two standard deviations above the mean is the same whatever the mean and standard deviation of the distribution. In fact, the proportion of values within two standard deviations of the mean is about 95.4% for any normal distribution.

Solution 5.4

In this case, all the areas between a and b are equal to 0.997 (to 3 s.f.).

The area under a normal curve from three standard deviations below the mean to three standard deviations above the mean is the same whatever the mean and standard deviation of the distribution. In fact, the proportion of values within three standard deviations of the mean is about 99.7% for any normal distribution.

Solution 5.5

(a) The area to the left of 1.644 85 is 0.95, so the area between $-1.644\,85$ and 1.644 85 is equal to 0.9. The required value is $z = 1.644\,85$ (to 6 s.f.).

(b) The area under the normal curve between $\mu - z\sigma = 20 - 1.644\,85 \times 5 = 11.775\,75$ and $\mu + z\sigma = 20 + 1.644\,85 \times 5 = 28.224\,25$ is equal to 0.9.

(c) Whatever values of μ and σ you choose, you should find that the area under the normal curve between $\mu - 1.644\,85\sigma$ and $\mu + 1.644\,85\sigma$ is equal to 0.9.

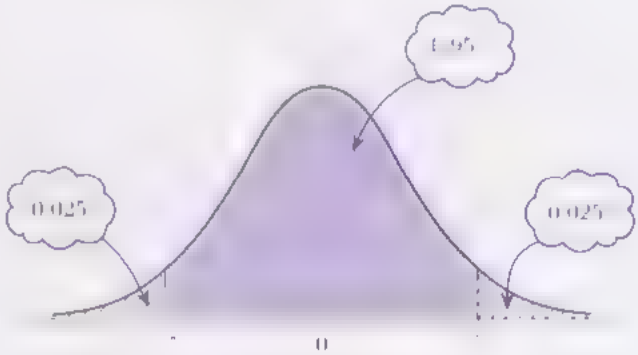
(d) The area under a normal curve within 1.644 85 standard deviations of the mean is 0.9, whatever the values of the mean and standard deviation

In practice, we shall normally use the value of z calculated to only 3 significant figures (which is 2 decimal places in this case). Rounding to 2 decimal places gives 1 64, and this is the value that is commonly used; we do not usually require greater accuracy than this. So we have the following result.

For a population modelled by a normal distribution with mean μ and standard deviation σ, approximately 90% of the population are within 1.64 standard deviations of the mean, that is, between $\mu - 1.64\sigma$ and $\mu + 1.64\sigma$.

Solution 5.6

(a) If the area between $-z$ and z is equal to 0.95, then the area of each tail is $\frac{1}{2} \times 0.05 = 0.025$. So the total area to the left of z is 0.975. This is illustrated in Figure S2.4.



(a) The area between $-z$ and z

(b) The area to the left of z

Figure S2.4 Finding the area to the left of z

The area to the left of 1.959 96 is equal to 0 975 The required value is $z = 1.959\,96$ (to 6 s.f).

(b) The area under the normal curve between $\mu - z\sigma = 20 - 1\,959\,96 \times 5 = 10\,2002$ and $\mu + z\sigma = 20 + 1.959\,96 \times 5 = 29.7998$ is equal to 0 95

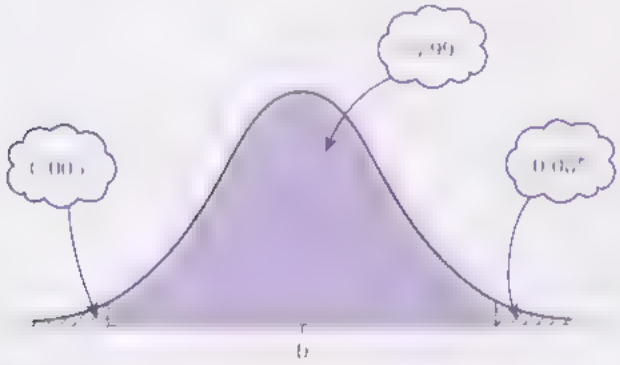
(c) Whatever values of μ and σ you choose, you should find that the area under the normal curve between $\mu - 1.959\,96\sigma$ and $\mu + 1.959\,96\sigma$ is equal to 0.95.

(d) The area under a normal curve within $1.959\,96 \simeq 1.96$ standard deviations of the mean is 0.95, whatever the values of the mean and standard deviation.

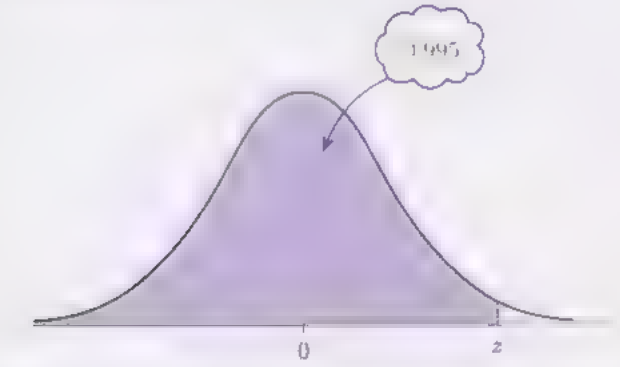
For a population modelled by a normal distribution with mean μ and standard deviation σ, approximately 95% of the population are within 1.96 standard deviations of the mean, that is, between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$.

Solution 5.7

(a) If the area between $-z$ and z is equal to 0.99, then the area of each tail is $\frac{1}{2} \times 0.01 = 0.005$. So the total area to the left of z is 0.995. This is illustrated in Figure S2.5.



(a) The area between $-z$ and z



(b) The area to the left of z

Figure S2.5 Finding the area to the left of z

For the area to the left of z to be 0.995, z must be equal to 2.575 83 (to 6 s.f.).

(b) The area under the normal curve between $\mu - z\sigma = 20 - 2.575\,83 \times 5 = 7.120\,85$ and $\mu + z\sigma = 20 + 2.575\,83 \times 5 = 32.879\,15$ is equal to 0.99

(c) Whatever values of μ and σ you choose, you should find that the area under the normal curve between $\mu - 2.575\,83\sigma$ and $\mu + 2.575\,83\sigma$ is equal to 0.99

(d) The area under a normal curve within $2.575\,83 \simeq 2.58$ standard deviations of the mean is 0.99, whatever the values of the mean and standard deviation.

For a population modelled by a normal distribution with mean μ and standard deviation σ, approximately 99% of the population are within 2.58 standard deviations of the mean, that is, between $\mu - 2.58\sigma$ and $\mu + 2.58\sigma$.

Chapter D3

Solution 3.5

The 95% confidence interval for the mean length (in millimetres) of cuckoo eggs given by *OUStats* is (22.2762, 22.5469). So, rounding to 3 significant figures, a 95% confidence interval for the mean length (in millimetres) of cuckoo eggs is (22.3, 22.5).

Solution 3.6

(a) Frequency diagrams of sentence lengths for each of the three authors are shown in Figure S3.1. In each case, a first interval starting value of 0.5 and an interval width of 5 were used.

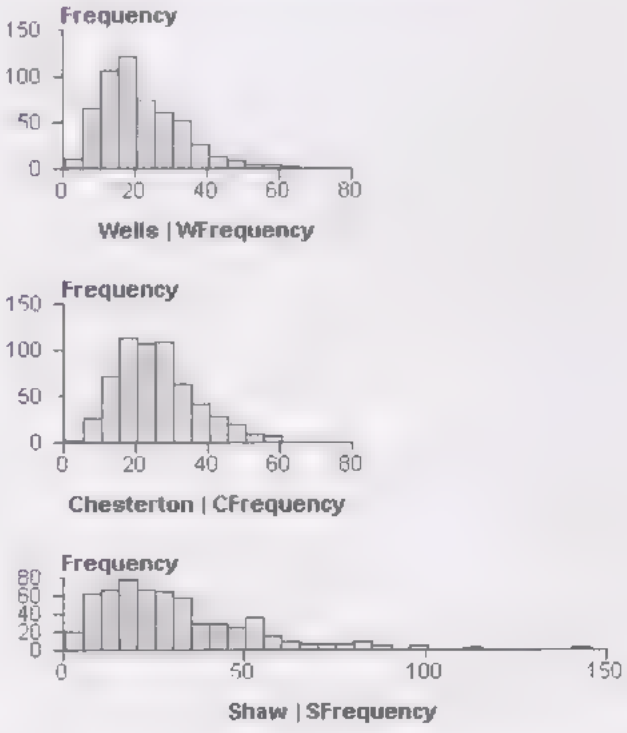


Figure S3.1 Frequency diagrams of sentence lengths

In order to produce these diagrams, which have different scales, I used **Tile** from the **Window** menu and then adjusted the sizes of the windows by dragging the mouse

The sentence lengths from the book by Shaw are much more variable than the sentence lengths from either of the other two books. There are more long sentences, and some very long ones.

(b) The mean sentence lengths for Wells, Chesterton and Shaw are (according to *OUStats*) 21.6799, 25.6131 and 31.1642, respectively. After rounding to 3 s.f., these are 21.7, 25.6 and 31.2. So, on average, Shaw seems to have written the longest sentences and Wells the shortest.

(c) The 95% confidence intervals for the mean sentence lengths in each of the three books, as given by *OUStats*, are as follows.

Wells	(20.7435, 22.6164)
Chesterton	(24.7502, 26.4759)
Shaw	(29.4516, 32.8767)

Rounding the confidence limits to 3 significant figures (one more than is given in the data) gives the following.

Wells	(20.7, 22.6)
Chesterton	(24.8, 26.5)
Shaw	(29.5, 32.9)

These confidence intervals are represented in the sketch in Figure S3.2.

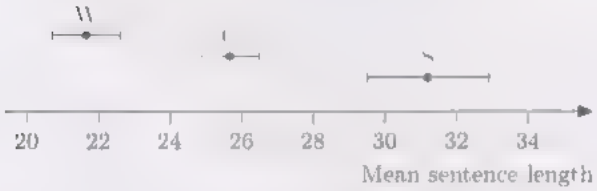


Figure S3.2 Confidence intervals

Since these confidence intervals do not overlap, this suggests that the mean sentence lengths in the books are different. It looks as though the mean sentence length in the book by Shaw is greater than the mean sentence length in the book by Chesterton, and that this is in turn greater than the mean sentence length in the book by Wells

However, we cannot say how confident we are that the means are different. For instance, although we are 95% confident that the mean length of sentences in the book by Wells is between 20.7 and 22.6, and we are 95% confident that the mean length of sentences in the book by Chesterton is between 24.8 and 26.5, we cannot put a figure to our confidence that the two means are different: this might be more or less than 95%, but we cannot say what it is simply by comparing the two 95% confidence intervals.

If we want to be able to quantify our confidence that the means are different, then a different approach is needed; a method which compares the two samples of data is required, rather than a method which looks at each sample separately. Such a method is discussed in the next chapter.

Solution 3.7

The 95% confidence intervals for the mean birthweights (in grams) of boys and girls born two weeks early, as given by *OUStats*, are as follows

Boys	(3068.36, 3349.53)
Girls	(2912.11, 3239.48)

These confidence intervals are represented in the sketch in Figure S3.3.

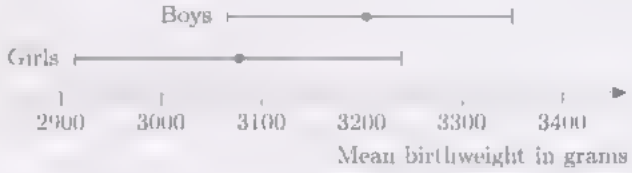


Figure S3.3 Confidence intervals

Although the confidence limits are higher for the boys than for the girls, the two intervals overlap, so we cannot draw conclusions from these confidence intervals about whether there is a difference between the mean birthweights of boys and girls born two weeks early.

Chapter D4

Solution 2.4

Boxplots of the men's and women's earnings are shown in Figure S4.1.

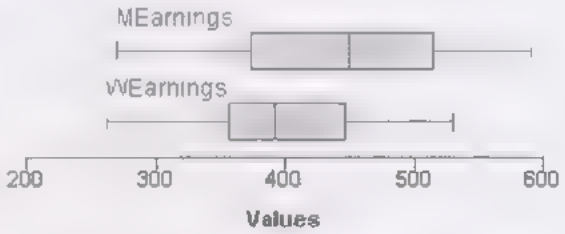


Figure S4.1 Gross weekly earnings (in pounds) of male and female primary school teachers

The earnings of the men seem to be generally higher than the earnings of the women, although the difference does not appear to be great. The difference is greatest for the highest earners in the two groups, and is very small for the lowest earners in the two groups.

Solution 4.3

(a) The boxplots are shown in Figure S4.2.

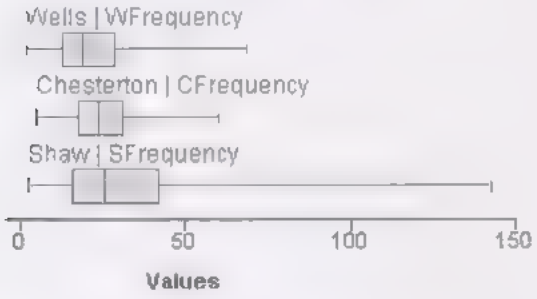


Figure S4.2 Boxplots of sentence lengths

As you can see, all the books contained quite a high proportion of fairly short sentences; and all three boxplots are right-skew, indicating that long sentences are less common than shorter ones in all three books. The main difference between the sentence lengths in the three samples seems to be that some of the sentences in the book by Shaw are longer than any of the sentences in the samples from the other two books. The average sentence length, as measured by the median, is longest for the sample from the book by Shaw and shortest for the sample from the book by Wells.

(b) The null and alternative hypotheses may be written as

$$H_0 : \mu_C = \mu_W,$$
$$H_1 : \mu_C \neq \mu_W,$$

where μ_C is the mean sentence length in the book by G. K. Chesterton, and μ_W is the mean sentence length in the book by H. G. Wells.

The test statistic (obtained from *OUStats*) is $z = 6.05118$.

Since the test statistic $z = 6.05418 > 1.96$, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that the mean sentence length in the book by Chesterton is not equal to the mean sentence length in the book by Wells. The sample mean is greater for the book by Chesterton than for the book by Wells, so this suggests that the mean sentence length is greater for the book by Chesterton than for the book by Wells.

(c) The null and alternative hypotheses may be written as

$$H_0 : \mu_C - \mu_S,$$
$$H_1 : \mu_C \neq \mu_S,$$

where μ_C is the mean sentence length in the book by G. K. Chesterton, and μ_S is the mean sentence length in the book by G. B. Shaw.

The test statistic (obtained from *OUStats*) is $z = -5.67378$.

Since the test statistic $z \quad -5.67378 < -1.96$, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that the mean sentence length in the book by Chesterton is not equal to the mean sentence length in the book by Shaw. The sample mean is greater for the book by Shaw than for the book by Chesterton, so this suggests that the mean sentence length is greater for the book by Shaw than for the book by Chesterton.

Solution 4.4

(a) The boxplots are shown in Figure S4.3.

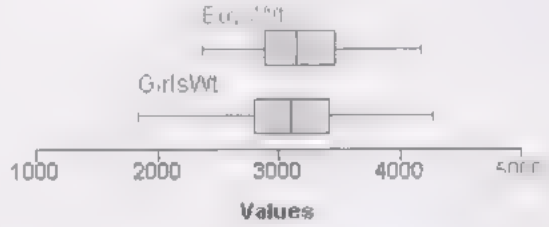


Figure S4.3 Boxplots of birthweights

No great difference is apparent between the birthweights of the boys and the girls, although the median birthweight is slightly higher for the boys than for the girls, and there is less spread in the boys' birthweights.

(b) The null and alternative hypotheses may be written as

$$H_0 : \mu_B = \mu_G,$$
$$H_1 : \mu_B \neq \mu_G,$$

where μ_B is the mean birthweight (in grams) of baby boys born two weeks early, and μ_G is the mean birthweight of baby girls born two weeks early.

The test statistic (obtained from *OUStats*) is $z = 1.20951$.

Since $-1.96 < z < 1.96$, we cannot reject the null hypothesis at the 5% significance level. The data do not provide sufficient evidence at the 5% significance level to reject the hypothesis that the mean birthweight of boys born two weeks early is equal to the mean birthweight of girls born two weeks early.

Solution 4.5

The null and alternative hypotheses may be written as

$$H_0 : \mu_M = \mu_F,$$
$$H_1 : \mu_M \neq \mu_F,$$

where μ_M is the mean gross weekly earnings (in pounds) of male primary school teachers in 1995, and μ_F is the mean gross weekly earnings of female primary school teachers in 1995.

The test statistic (obtained from *OUStats*) is $z = 2.79034$.

Since the test statistic $z = 2.79034 > 1.96$, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that there was a difference between the mean gross weekly earnings in 1995 of male and female primary school teachers. The sample mean is greater for the men than for the women, so this suggests that the mean gross weekly earnings of male primary school teachers in 1995 was greater than the mean gross weekly earnings of female primary school teachers in 1995.

Solution 6.2

(a) The scatterplot and the least squares fit line are shown in Figure S4.4. Note that each 'plus' on the scatterplot may represent the heights of one father–son pair or of many pairs: it is not possible to tell how many, as frequencies are not represented. If you feel that the line shown does not look as if it is the best fit line, then this is probably because when looking at the scatterplot you cannot take into account the relative frequencies of the different pairs of values.

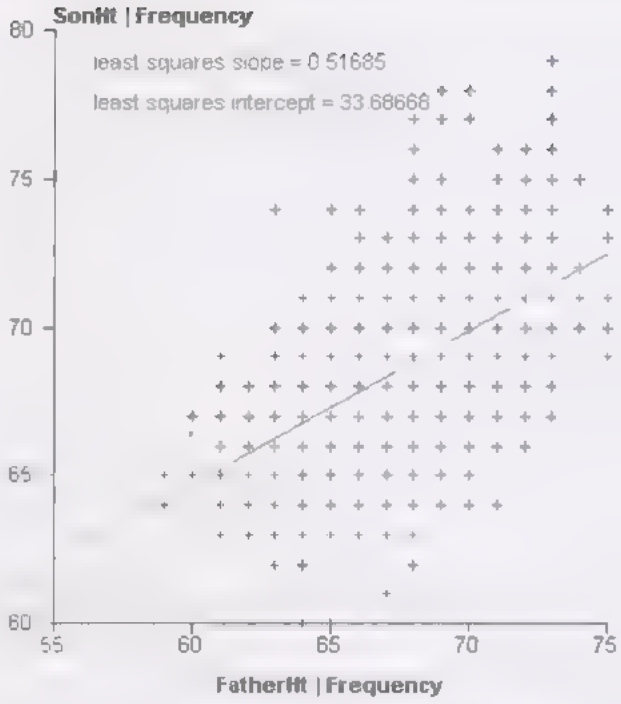


Figure S4.4 A scatterplot of son's height against father's height

(b) The equation of the least squares fit line is

$$y = 33.69 + 0.5169x,$$

where y represents son's height and x represents father's height (giving the parameters to 4 s.f.).

The predicted height of the son of a 70-inch-tall man is

$$y = 33.69 + 0.5169 \times 70 \simeq 69.9 \text{ inches.}$$

However, the data on which the model is based were collected in the 1890s for fathers and sons in the UK. If the average height of men has continued to increase from generation to generation, possibly by different amounts in different generations, then data collected now might well lead to a slightly different model. This prediction applies only to the sons of fathers in the UK in the 1890s. Moreover, the families measured in Pearson's study were predominantly middle class, so the prediction applies to middle class families in the 1890s. (A different model might have been required for the heights of fathers and sons in working class families.)

(c) There is a lot of scatter in the plot, so any individual son of a 70-inch-tall man could be a lot taller or shorter than 69.9 inches. This height is an estimate of the mean height of sons of 70-inch-tall men in the UK in the 1890s.

Solution 6.3

(a) The scatterplot is shown in Figure S4.5.

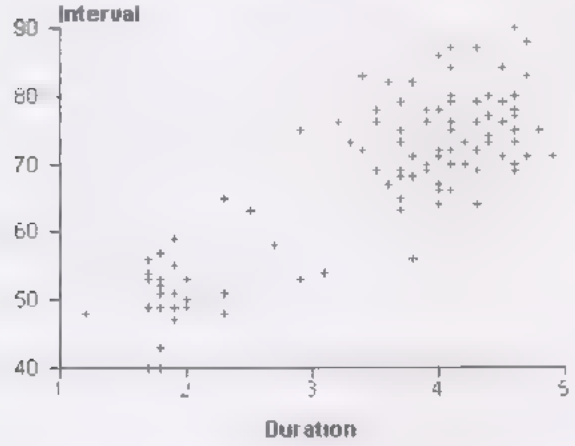


Figure S4.5 A scatterplot of time to next eruption against duration of eruption

(b) There seem to be two main groups of points on the scatterplot, corresponding to 'short' and 'long' eruptions. 'Short' eruptions lasted two minutes or less; 'long' eruptions lasted more than three minutes. Very few of the eruptions were of intermediate duration. However, overall, the time to the next eruption appears to increase with the length of the current eruption. Following a long eruption, there is a longer wait, on average, until the the next eruption than following a short eruption. However, there is a lot of scatter in the plot, so the relationship is not a very strong one. It does look as though a straight line might summarise the relationship quite well.

(c) The equation of the least squares fit line is

$$y = 33.65 + 9.806x,$$

where x minutes is the duration of an eruption and y minutes is the waiting time until the start of the next eruption.

(i) After an eruption of length 1.5 minutes, the predicted time until the start of the next eruption is

$$33.65 + 9.806 \times 1.5 \simeq 48.4 \text{ minutes.}$$

(ii) After an eruption of length 3 minutes, the predicted time until the start of the next eruption is

$$33.65 + 9.806 \times 3 \simeq 63.1 \text{ minutes.}$$

(iii) After an eruption of length 4.5 minutes, the predicted time until the start of the next eruption is

$$33.65 + 9.806 \times 4.5 \sim 77.8 \text{ minutes.}$$

(d) The predictions are estimates of the *mean* waiting time until the next eruption following eruptions of lengths 1.5, 3 and 4.5 minutes, and there is a lot of scatter about the regression line. So the next eruption may be much sooner or much later than predicted. Nevertheless, the predictions could be used to give a very rough indication of when the next eruption is likely to occur: the mean waiting time is nearly half an hour longer following a 4.5-minute eruption than following a 1.5-minute eruption. This is a situation where a confidence interval might be more useful than a simple prediction. The lower confidence limit might be useful as an indication of the earliest time that the next eruption is likely to occur.

Solution 6.4

(a) There is a lot of scatter in the plot for the elderly group, but it does look as though there might be a weak relationship between memorisation time and city block score. The equation of the least squares fit line is

$$y = 36.81 - 0.1793x,$$

where x is the memorisation time in seconds, and y is the city block score.

(b) Again there is a lot of scatter in the plot for the young group, but less than in the scatterplot for the elderly group. The equation of the least squares fit line is

$$y = 32.61 - 0.1836x.$$

(c) The equation of the least squares fit line for the combined group is

$$y = 38.07 - 0.2264x.$$

According to this model, the predicted city block scores are as follows

(i) When the memorisation time is 1 minute, the predicted score is

$$38.07 - 0.2264 \times 60 \simeq 24.$$

(ii) When the memorisation time is 2 minutes, the predicted score is

$$38.07 - 0.2264 \times 120 \simeq 11.$$

(iii) When the memorisation time is 3 minutes, the predicted score is

$$38.07 \quad 0.2264 \times 180 \simeq \quad .3$$

Clearly, a city block score of −3 is impossible: the lowest possible score is 0, for someone who replaces all the objects in the correct positions. Note that three minutes is outside the range of memorisation times for the people taking the test, so using the model for prediction is not valid in this case. Clearly, the model is not appropriate for times as long as three minutes. Since the lowest possible score is 0, a model which does not predict values below zero might be considered. Might a curve rather than a straight line provide a more useful model?

Solution 6.5

In order to draw the lines, I found the coordinates of two points on each line: (0, 36.8) and (150, 9.9) for the fit line for the elderly group; (0, 32.6) and (150, 5.1) for the fit line for the young group. The two least squares fit lines are drawn on the scatterplot in Figure S4.6.

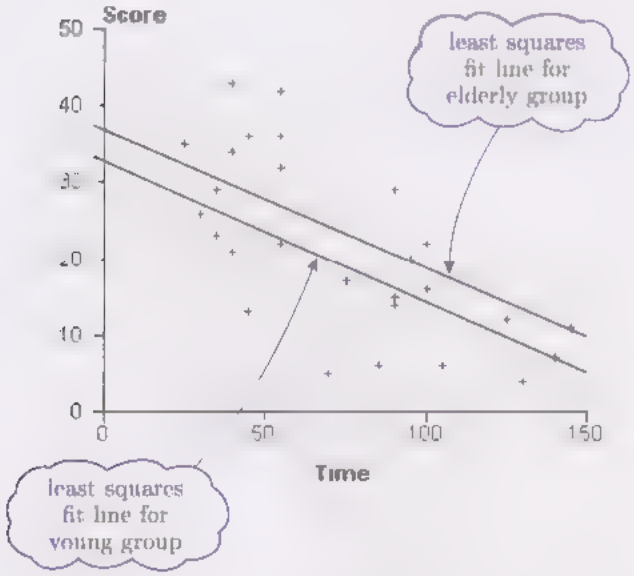


Figure S4.6 The two least squares fit lines

As you can see, the gradients of the two lines are roughly equal. However, the line for the elderly group is a little higher than the line for the young group, suggesting that elderly people do not perform quite as well as young people who spend similar times memorising the positions of the objects. Is there a real difference, or is the observed difference between the lines simply due to sampling variation? This is the sort of question that more advanced regression techniques can tackle. It is possible to fit parallel lines to two data sets and then to carry out a hypothesis test of whether there is a 'real' difference between the intercepts or whether the observed difference might be due to chance. However, we shall not be discussing how to do so in this course.

Index for OUStats

area under a normal curve, 19

boxplot, 47, 49

calculator, 18
changing the size of a window, 27
closing a window, 21
confidence interval, 44, 46
copying from a window, 21
creating a new data file, 57

data file, 18, 57
data matrix, 57
data window, 14, 57

editing data, 57
entering data, 57
Exit, 14

file name, 15, 18
file name extension, 15
first interval starting value, 19
fitting a normal curve, 19, 25
frequency data, 21, 58
frequency diagram, 19, 22

histogram, 19

interval width, 19

key values on a boxplot, 47, 49

least squares fit line, 53, 56

menus for *OUStats*, 15, 18

naming a column, 57
normal distribution, 19, 38
normal samples, 25, 28
Notes file, 18, 58

opening a data file, 18
opening window in *OUStats*, 14

plotting symbol, 56
printing, 29

random sample, generating, 25, 28
regression line, 53, 56
removing regression line, 53
renaming a column, 57

sample size too small, 45, 51
saving a data file, 21
saving a new data file, 57
scatterplot, 32, 37, 56
selecting several variables, 18, 25
summary statistics, 18

Tile, 26
transferring data into and out of *OUStats*, 58
two-sample z-test, 50, 52

Acknowledgements

Grateful acknowledgement is made to the following sources for permission to reproduce material in this book and on the course CD.

Draper, N. R. and Smith, H. (1966), *Applied Regression Analysis*, reprinted by permission of John Wiley and Sons, Inc. All rights reserved.

Fowler, J. and Cohen, L. (1996), *Statistics for Ornithologists, BTO Guide 22*, Second Edition, British Trust for Ornithology.

Reprinted with permission from Johnson, M. P. and Raven, P. H. (1973), 'Species number and endemism: the Galápagos Archipelago revisited', *Science*, **179**, pp. 893–895, Copyright © 1973 American Association for the Advancement of Science.

Lindgren, B. W. and Berry, D. A. (1981), *Elementary Statistics*, Macmillan Publishing Co., Inc., by permission of Prentice-Hall, Inc.

Mazess, R. B., Peppler, W. W. and Gibbons, M. (1984), 'Total body composition by dual-photon (^{153}Gd) absorptiometry', *American Journal of Clinical Nutrition*, **40**, pp. 834–839.

Metzger, W. H. (1935), *Journal of the American Society of Agronomy*, **27**, p. 653, American Society of Agronomy.

Selvin, S. (1991), *Statistical Analysis of Epidemiological Data*, New York, Oxford University Press.

Snedecor, G. W. and Cochran, W. G. (1980), *Statistical Methods*, Seventh Edition, The Iowa State University Press.

Sokal, R. R. and Rohlf, F. J. (1981), *Biometry*, Second Edition, W. H. Freeman and Company Publishers.

Spiegel, M. R. (1972), *Schaum's Outline of Theory and Problems of Statistics in SI Units*, reproduced with permission of The McGraw-Hill Companies.

Terman, L. M. (1919), *The Intelligence of School Children*, Copyright © 1919 by Houghton Mifflin Company. Renewed 1946. Reprinted with the permission of the author's estate.

Trumple, R. J. and Weaver, H. F. (1953), *Statistical Astronomy*, p. 194, Table 1.5, Copyright © 1953 The Regents of the University of California.

Wetherill, G. B. (1981), *Intermediate Statistical Methods*, Chapman and Hall.

The Open University
ISBN 0 7492 0275 0